



Interactive clustering: a scoping review

Thais Rodrigues Neubauer¹ · Sarajane Marques Peres¹ · Marcelo Fantinato¹ · Xixi Lu² · Hajo Alexander Reijers²

Published online: 8 October 2020
© Springer Nature B.V. 2020

Abstract

We present in this paper a scoping review conducted in the interactive clustering area. Interactive clustering has been applied to leverage the strengths of both unsupervised and supervised learning. In interactive clustering, supervised learning is represented by inserting the knowledge of human experts in an originally unsupervised data analysis process. This scoping review aimed to organize the knowledge on (i) the applicability of interactive clustering methods, (ii) clustering algorithms being used to support interactive clustering, (iii) how to model the expert supervision and (iv) the effects brought by the expert supervision in the results produced. A systematic search for related literature was conducted in the Scopus database, resulting in the selection of 50 primary studies published by 2018. The analysis of these studies allowed us to identify trends such as: the application in text/image; use of partitioning and hierarchical algorithms; application of strategies based on split/merge, pairwise constraints, similarity metrics learning and data reassignment; and concern with visualization. In addition, some relevant issues not yet adequately addressed were identified, such as: the evaluation of expert supervision; the evaluation of the expert's effort; and the conduction of studies effectively involving human experts, instead of computer simulations.

Keywords Interactive clustering · Active learning · Human-in-the-loop · Clustering · Expert supervision · User supervision

✉ Thais Rodrigues Neubauer
thais.neubauer@usp.br

Sarajane Marques Peres
sarajane@usp.br

Marcelo Fantinato
m.fantinato@usp.br

Xixi Lu
x.lu@uu.nl

Hajo Alexander Reijers
h.a.reijers@uu.nl

¹ University of São Paulo, São Paulo, Brazil

² Utrecht University, Utrecht, The Netherlands

1 Introduction

Interactive clustering is a data analysis approach that includes a human expert in key decisions of the clustering process (Hu et al. 2014; Schwenker and Trentin 2014). Including an expert in the loop of unsupervised data analysis aims to achieve higher quality results or results that are aligned with specific needs of a particular user or scenario. The difficulty in obtaining quality or compliance in data analysis results in the unsupervised analysis context is usually due to the assumptions or arbitrary decisions made during the parameterization of the clustering algorithms. These assumptions and arbitrary decisions may not correspond to the actual distributions of data under analysis. In addition, they rarely map all the relevant information that the data represent in the scope of the real problem on which they are generated (Lei et al. 2017).

Although the term *interactive clustering* has only recently become more popular, one of the first scientific works to address the systematic insertion of human knowledge interactively in an originally unsupervised data analysis process dates back to 1971. The pioneering approach of Patrick and Shen (1971) inserts in the clustering context the expert's knowledge on the data, in the form of *mean vectors*, *covariance matrices* and *expert's confidences* that accurately characterize the data distribution. The interactive nature of their approach is strengthened by: (a) expert's knowledge is continually used by the clustering algorithm whilst new data points are inserted into the clustering process; (b) expert's knowledge influences the extent to which a new data point should change the current result of the clustering process; and (c) the experts themselves learn about the problem and adjust their beliefs by observing the actions taken by the clustering process.

Patrick and Shen (1971) faithfully represent the understanding of interactive clustering adopted in our study. Figure 1 illustrates this view. Any expert interaction during pre- or post-processing tasks is considered part of a broader interactive data analysis process, which is beyond the scope of this study.

Interactive clustering is costly for involving humans, and modeling this type of solution is complex and challenging because the following reasons. Firstly, clustering algorithms are sensitive to the size and complexity of the problem under resolution. Involving humans requires clustering algorithms to be tailored to accept human intervention and assumes

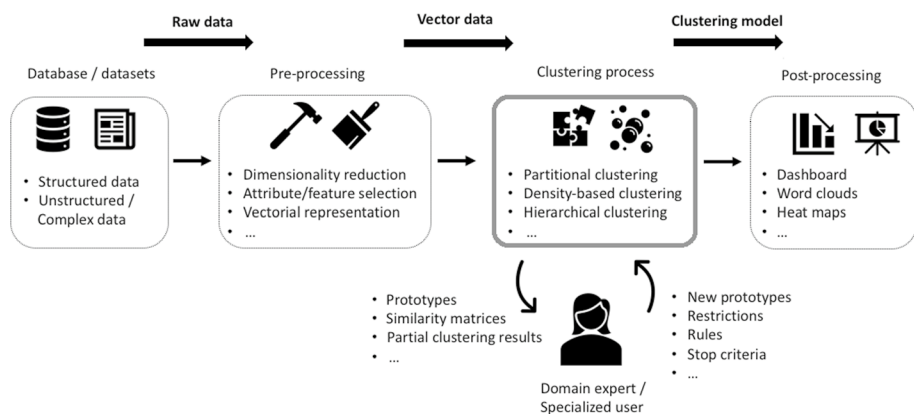


Fig. 1 Interactive clustering approach: the domain expert or specialized user works in conjunction with the clustering algorithm during the clustering process

that the algorithms' outputs are human-interpretable. Moreover, the effects of human intervention should be assessed for the quality of clustering results, and the effort spent by the expert should lead to a positive cost/benefit relationship regarding the quality of the achieved solutions. In addition, obtaining and analyzing knowledge on interactive clustering involves treating information under different taxonomies as similar or equivalent, since approaches appear with different nomenclatures in the specialized literature. These aspects show that there are many challenges associated with the research and practice of interactive clustering. These challenges motivate the presentation of this review as a facilitator for researchers interested in starting studies or positioning their studies in relation to what has been developed in the area.

This paper presents a scoping review that organizes information on the extent, range and nature of research activities in the interactive clustering area as well as identifies open issues that require attention from researchers and practitioners interested in this area. Following Paré et al. (2015)'s guidelines, this scoping review aims to provide extensive and comprehensive coverage of the interactive clustering literature and its trends, guided by answering the following broad scope research questions:

- RQ-1** In what data types and application domains has interactive clustering been applied to?
- RQ-2** Which classes of clustering algorithms have been used in interactive clustering? Why were they chosen?
- RQ-3** How has expert supervision been implemented in interactive clustering?
- RQ-4** How has expert supervision been evaluated considering the clustering outcomes? What effects does this expert supervision have on the clustering process and in the clustering outcomes?
- RQ-5** How has the expert's effort been measured?

This scoping review is organized as follows: Sect. 2 introduces the basic concepts necessary for understanding the further discussions; Sect. 3 positions this scoping review within related systematization efforts for the areas similar to interactive clustering area; Sect. 4 details the research questions that guide the acquisition of information and the systematization of the knowledge obtained; Sect. 5 presents the research method followed in this scoping review and details of the conduction process; Sect. 6 organizes the contributions of this scope review through the findings and discussions; and finally, Sect. 7 presents an analysis of validity threats and our concluding remarks.

2 Background

In this section, we present basic concepts of clustering approaches. These concepts concern two important machine learning paradigms (supervised and unsupervised) and the modeling of both the classic clustering and the interactive clustering tasks.

2.1 Supervised learning and unsupervised learning

In machine learning, the focus is on inductive learning, which occurs based on observations, represented in a dataset related to the phenomenon about which one intends to learn. According to Mitchell (1997), a computer program learns if it can improve its performance

in a given task, based on previous experiences, under some measure of performance evaluation. Performance improvement occurs through the execution of an algorithm whose aim is to optimize the performance measure. The way this measure is optimized characterizes the learning paradigm followed by the algorithm:

- *Supervised learning*: when the algorithm performance measure refers to an error between the output produced by the algorithm and the expected previously known output. In this paradigm, given a collection of data $(x, f(x))$, it is expected that the optimization process will produce a function h that approximates f .
- *Unsupervised learning*: when the algorithm performance measure is based on either the data similarity relationships or the data association probabilities assumed by the algorithm. In this paradigm, given a data collection x , it is expected that the optimization process will produce a relationship $R(x_i, x_j), \forall ij$, that explains behaviors underlying the phenomenon that generated such data collection.

2.2 Clustering and interactive clustering

Clustering is a descriptive task because it allows discovering profiles (i.e., behaviors) underlying a dataset and hence is implemented following the precepts of unsupervised learning (Fayyad et al. 1996).

The clustering task receives as input a dataset represented by a matrix $X \in \mathbb{R}^{n \times m}$. Each of its n rows represents a data point, and each of its m columns represents a feature that characterizes the data points. Matrix X is formed by a set of row vectors $\mathcal{N} = \{\vec{x}_1, \dots, \vec{x}_n\}$. The aim of the clustering task is to find k clusters in \mathcal{N} , denoted by subsets $\mathcal{K}_p \subseteq \mathcal{N}$, with $p \in \{1, \dots, k\}$, on which a performance measure is optimized. Thus, set $\mathcal{K} = \{\mathcal{K}_1, \dots, \mathcal{K}_k\}$ is a solution for the clustering task. An unsupervised algorithm implementing a solution for the clustering task adjusts the parameters W of a function \mathcal{G} based on the optimization of the performance measure. Therefore, for the classic clustering problem, the function:

$$\mathcal{G} : \mathbb{R}^{n \times m} \times W \rightarrow \mathcal{K} \quad (1)$$

maps the input data in matrix X to the subsets of row vectors of \mathcal{K} , such that:

1. $\mathcal{K}_p \neq \emptyset$, for all $p \in \{1, \dots, k\}$;
2. $\bigcup_{1 \leq p \leq k} \mathcal{K}_p = X$;
3. $\mathcal{K}_p \cap \mathcal{K}_q = \emptyset$, for all $p, q \in \{1, \dots, k\}$ and $p \neq q$.

Performance measurement is a key issue in this optimization process. To meet the profile discovery goal, a similarity function must be adopted based on either distance measurements or density criteria for data space occupation or joint probability information (Ester et al. 1996). The similarity function adopted characterizes the class of the clustering algorithm (Han et al. 2011) and represents the hypothesis that associates a data with a cluster (i.e., an element of \mathcal{K}). For example, when either the Euclidean distance or the cosine similarity is adopted, different assumptions are being made regarding the behavior of the data. Figure 2a illustrates a clustering resulted from the use of Euclidean distance while Fig. 2b illustrates the final clustering of the same dataset resulted from the use of cosine similarity.

In addition, other issues may also influence the results of clustering. For example, variations in the execution conditions of the algorithm (e.g., hyperparameter initialization, data sampling and data input strategy) can lead to different clustering results

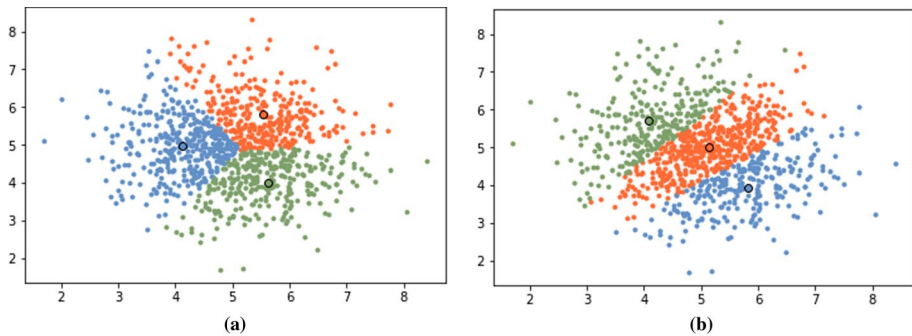


Fig. 2 Clustering resulting from the application of the *k-means++* algorithm, with $k = 3$ and using: **a** Euclidean distance; **b** cosine similarity. The obtained clusters are represented by colors. (Color figure online)

(Awasthi et al. 2017; Hu et al. 2014). The chosen data features as well as the adopted vector representation determine the organization of the data in the decision space, in which the similarity function is applied. Hence, these two elements also determines the profiles that can be found by the algorithm.

In order to minimize ambiguities resulting from arbitrary choices for the issues discussed above (i.e., for the performance measure, algorithm execution conditions, data features and the data vector representation), classic clustering approaches consider assumptions about the phenomenon that generates the data or arising from preliminary statistical analysis of the data (Awasthi et al. 2017; Rinaldo 2010; Achlioptas and McSherry 2005; Brubaker and Vempala 2008; Kalai et al. 2010; Moitra and Valiant 2010; Belkin and Sinha 2010; Chaudhuri and Dasgupta 2010). However, these classic clustering approaches do not verify the reasonableness of such assumptions for the actual context from which the data is derived. These assumptions hardly solve problems related to the ill-representation of important data features in the actual context (Coden et al. 2017; Lei et al. 2017; Vikram and Dasgupta 2016; Correa et al. 2015), which results from the limitations of adopted performance measures, algorithm execution conditions, data features and data vector representations.

Interactive clustering is an alternative to further maximize the chances of getting satisfactory clustering results. In interactive clustering, the knowledge of an expert in the actual context of the data is added in a controlled manner, aiming at contributing to the production of results more appropriate to the clustering specific application domain. In this approach, human judgments, knowledge, and expectations are combined with clustering algorithms (Chang et al. 2016; Hu et al. 2014).

Function (2) adapts Function (1) to include the influence of the human expert in the clustering process:

$$\mathcal{G}_I : \mathbb{R}^{n \times m} \times \oplus \times W \rightarrow \mathcal{K} \quad (2)$$

In this function, X , W and \mathcal{K} are as defined previously, and \oplus represents the expert's knowledge universe added in the clustering process. The expert's knowledge is acquired from an initial clustering provided by an optimization algorithm for the Function (1).

Implementing the interactive clustering task includes executing a set $A = \{a_n \mid n > 0\}$ of expert's actions, arising from the observation of \mathcal{K}_t , and imposing a set of restrictions¹ $R = \{r_m \mid m \geq 0\}$, arising from the expert's actions, on the process of obtaining the next set of clusters \mathcal{K}_{t+1} . If the expert's actions do not generate any restrictions ($m = 0$), then \mathcal{K}_{t+1} is produced without the expert's influence. The interactive clustering process ends with the expert signaling that the clustering result is satisfactory, getting an adequate value for the performance measure or reaching a maximum time or cost limit allowed for the process.

3 Related work

Table 1 summarizes seven reviews related to the study presented in this paper. No previously published review was found that specifically addresses the context of *interactive clustering*, as the one presented in this paper. The published reviews address broader contexts than interactive cluster, such as *partially supervised learning*, *active machine learning*² and *interactive data mining*.

Although the previously published reviews do not directly address our specific interest in interactive aspects for the clustering task, they partially cover it as they discuss approaches that combine human expert's knowledge with: machine learning, including some aspects related to the unsupervised learning paradigm, implementing clustering; or the automated process of discovering knowledge and implementing the entire data mining workflow. However, most of the related work includes analyzing very few studies focused on clustering, what happens even with studies that reviewed about 100 papers in total. Moreover, three of the published reviews focus on specific domains (Natural Language Processing (NLP), bioinformatics, and building effective Human-Computer Interface (HCI) designed for interactive machine learning) while we do not focus on any specific domain in order to deliver a more comprehensive view of the area of interactive clustering.

Finally, unlike our scope review presented here, none of the authors of these related reviews followed a systematic protocol for the search and selection of the primary studies to be considered in the review of the area. Only Dudley and Kristensson (2018) can be said to have followed a semi-systematic search protocol. By following a systematic protocol, we are able to present a comprehensive overview of the studies published in our area of interest. In addition, the first two related works are published as technical reports and are not peer-reviewed.

4 Research questions

The research questions were defined aiming to reflect on the interactive clustering area and thus produce a systematized knowledge about how it has been developing and what are its open issues. The defined research questions permeate the two main components

¹ Imposing restrictions on the clustering algorithm is implemented following the constrained clustering principles (Han et al. 2011; Chapelle et al. 2010). According to Wagstaff (2010), constrained clustering algorithms may enforce every constraint in the solution, or they may use the constraints as guidance rather than hard requirements.

² The term *active learning* is used when the learner has some role in determining on what data it will be trained (Cohn 2010), which can be done based on expert's knowledge.

Table 1 Comparison of related work [Document type: Technical Report (TR), Conference Paper (CP) and Journal Paper (JP)]

ID	Author (Year)	Document type	Context	Domain	Systematic?	Screened papers (#)	Interactive clustering papers (#)
1	Settles (2009)	TR	Active learning	—	—	≈ 150	7
2	Olsson (2009)	TR	Active learning	NLP	—	≈ 90	3
3	Sun and Wang (2010)	CP	Active learning	—	—	≈ 35	1
4	Gharehchopogh (2010)	CP	Interactive data mining	—	—	≈ 35	3
5	Holzinger et al. (2014)	JP	Interactive data mining	Bioinf.	—	≈ 70	0
6	Schwenker and Trentin (2014)	JP	Partially superv. learning	—	—	≈ 115	13
7	Dudley and Kristensson (2018)	JP	Interactive learning	HCI	±	≈ 100	6
8	Our study	JP	Interactive clustering	—	+	50	50

of interactive clustering: clustering algorithms and expert interactions during the application of such algorithms.

RQ-1 In what data types and application domains has interactive clustering been applied to?

When the nature of the data under analysis provides for the successful application of a class of algorithms or strategies for analyzing results (e.g., image data motivates applying algorithms that deal with arbitrary forms of clusters and visualization-based strategies for analysis of results).

The domain in which interactive clustering is applied can affect the choice of the problem-solving strategy. This occurs, for example, in the following situations: (a) when the nature of the data under analysis enables the successful application of a class of algorithms or strategies for analyzing results (e.g., image data motivates applying algorithms that deal with arbitrary forms of clusters and visualization-based strategies for analysis of results); or (b) when the size of the problems requires levels of efficiency met by a class of algorithms or strategies for analyzing results (e.g., if efficiency is a requirement, an instance-based constrained clustering algorithm is prohibitive and hence the expert should analyze the resulting clusters through representations with high levels of abstraction).

The answer to this research question aims to build an overview of the applicability of interactive clustering. This overview should be based on: (a) the type of the data involved in the interactive clustering approaches found and whether these initiatives are restricted to these data types or to specific datasets and (b) the application domains in which interactive clustering has been explored. This classification should increase the possibility of expanding the use of interactive clustering as well as pointing out the current limitations in implementation strategies.

RQ-2 Which classes of clustering algorithms have been used in interactive clustering? And why were they chosen?

As part of the solution design required to apply the clustering task to problem-solving, choosing which clustering algorithm to use is a key decision. In interactive clustering, the choice of a clustering algorithm should consider its suitability to be influenced by external decisions. Inserting an expert in the clustering task imposes constraints on associating data with clusters, and the algorithmic strategy adopted should be able to accept such constraints.

The answer to this research question should allow us to organize the information on: (a) which classes of clustering algorithms have been considered appropriate for interactive clustering; (b) why these classes of clustering algorithms have been chosen over others; and (c) whether the reasons identified justify the presence and absence of certain classes of algorithms.

To guide the search for information to answer this research question, the taxonomy of clustering algorithms proposed by Han et al. (2011) is adopted. This taxonomy comprises four classes: *partitioning*, *model-based*, *density-based* and *hierarchical methods*. In addition, *other* was adopted for cases where none of the previous classes were identified.

RQ-3 How has expert supervision been implemented in interactive clustering?

Expert supervision can be incorporated into the clustering process in a variety of ways. For example, the selection and presentation of information to be submitted to expert supervision may be done in different ways (e.g., visualization of data points or graphs with the characterization of clusters). In addition, how the expert's knowledge is inserted into the clustering process may also vary (e.g., using pairwise constraints or feature weighting).

To answer this research question, the ways of expert supervision were studied under four aspects: (a) how the expert's knowledge is acquired, i.e., split/merge requests, seeds indication or pairwise constraints (must-link/cannot-link); (b) how supervised information is inserted into the clustering process, i.e., in a similarity metric learning, executing splits/merges or using received pairwise constraints; (c) strategies for presenting elements for analysis to the expert and whether resources are available to support expert interaction, usually some computational tool that visually represents the information to be supervised; and (d) interaction synchronization, which can be done online, in which the expert is exclusively dedicated to the supervision task, or offline, in which interaction occurs asynchronously, with requests that require extra expert actions (such as searching information).

RQ-4 How has expert supervision been evaluated considering the clustering outcomes? And what effects does this expert supervision have on the clustering process and in the clustering outcomes?

The main goal of interactive clustering is to improve clustering results in some aspect. Thus, evaluating clustering results obtained with expert supervision allows us to understand the effects that supervision has on these results. Moreover, expert supervision can introduce adverse conditions to clustering algorithms, leading to the need to implement workaround procedures and evaluate the increase in the cost of executing the algorithms.

To answer this research question, an analysis was carried out on what evaluation methods researchers have been applying in their studies on interactive clustering and how the results of evaluations have been used to motivate or demotivate the inclusion of experts. From a clustering quality standpoint, we considered the application of internal and external clustering validation indices (Halkidi et al. 2001), applied to results obtained with and without expert supervision. From the standpoint of expert supervision usefulness, we considered the application of evaluations showing the suitability of the results to the context of the problem under analysis, even if such results do not represent quality clustering from the standpoint of validation indices above mentioned.

RQ-5 How has the expert's effort been measured?

Only good results obtained with expert supervision in the clustering process do not justify the use of interactive clustering. The effort required by the expert needs to deliver a positive cost-benefit ratio. This research question aims to discuss how the expert's effort has been measured. The answer to this question is explored under the following aspects (Marrero and Urbano 2018): time spent by the expert to perform supervision, ratio of amount and complexity of information submitted to the expert, amount of information accepted for the expert supervision process and amount of actions required during expert supervision.

Finally, some approaches aim to create computational agents to simulate the action of human experts in the clustering process. This simulation practice should be considered

Table 2 Canonical search string

```
(((interactive* OR "human in the loop*" OR "human computation" OR "active learn*") OR ((human*
OR user* OR client* OR customer* OR expert*) W/3 (collaborat* OR cooperat* OR supervis* OR
knowledge OR participat* OR interact*))) W/3 (cluster* OR unsupervis*)) AND (cluster* OR mining
OR "machine learning" OR "data science" OR "pattern recogni*" OR intelligen*))
```

in the discussion of this research question as there are distinct premises when human experts or computational agents are used in supervision.

5 Research method

In this section, we describe the protocol followed in this scoping review. The protocol was designed aiming at maximizing the reproducibility of the review and minimizing bias in the results produced, based on the prerogatives sought for a systematic review (Kitchenham and Charters 2007; Biolchini et al. 2005). This section is divided into two subsections: the first presents the procedures for conducting the scoping review and the second shows false positive examples of papers that are not considered primary studies for this scoping review.

5.1 Review protocol

The protocol was designed to systematize and direct actions to select primary studies to achieve answers to research questions. The steps of this protocol are: (A) selection of databases, (B) elaboration of search strategy, (C) definition of selection criteria for primary studies, (D) definition of the method for data extraction and results synthesis and (E) definition of quality criteria for primary studies.

- A. *Databases* We are interested in interactive clustering from a computer science perspective. A number of publishers publish papers with this focus, including Springer, ACM, IEEE, Elsevier, Taylor & Francis, Inderscience, Emerald, World Scientific and Wiley. Papers published by these publishers and others are indexed by Scopus³, well known as the largest database of peer-reviewed literature. Thus, Scopus was chosen to systematically explore the databases of these publishers and other related ones. Due to the large number of databases indexed by Scopus, this scoping review also covered university publishing houses and other scientific associations. Direct searches in the IEEE, ACM or Springer libraries, for example, would be redundant and hence unnecessary.
- B. *Search strategy* A search string was created comprising keywords related to interactive clustering. Through exploratory analysis, we found that researchers have used different terms to refer to interactive clustering. Thus, the search string comprises disjunctions of synonymous terms with interactive clustering obtained in the specialized literature. In addition, the search string has a second part designed to filter only by papers specifically related to data mining or related areas. A canonical representation for the search

³ <https://www.elsevier.com/solutions/scopus/how-scopus-works/content>.

Table 3 Inclusion Criteria (IC)

IC-1	The main research goal or one of the specific research goals of the study is interactive clustering, taking the definition of interactive clustering in Sect. 2
IC-2	The study addresses the application of clustering computational techniques with user intervention and discusses the results of its application
IC-3	User intervention occurs during cluster creation leading to results achieved necessarily due to human intervention

Table 4 Exclusion Criteria (EC)

EC-1	The study is published before 2007
EC-2	The study is not fully in English
EC-3	The study is not primarily related to computer science, information systems, engineering or strongly related field
EC-4	The study is not a peer-reviewed scientific paper
EC-5	The study is not primary
EC-6	The study is a preliminary version related to a later and more relevant publication

string is given in Table 2. To limit in a controlled way the breadth of the results brought by the first part of the search string, we used the proximity operator *W/n*, which defines the distance between words, but not the order. For example, *user W/3 cluster* defines that *user* can be found within three words of *cluster* in any order.

- C. *Selection criteria for primary studies* Since false-positive results could be returned due to the broad spectrum of the search string, these results needed to be verified. Tables 3 and 4 show, respectively, the Inclusion Criteria (IC) and the Exclusion Criteria (EC) defined for this verification. For a returned study to be selected as a primary study, it should meet all inclusion criteria and not any exclusion criteria.
- D. *Data extraction and results synthesis* The papers selected for inclusion were read in full for extraction of results. Reading was guided by the research questions (cf. Sect. 4). While reading, papers' excerpts helping to answering the research questions were copied to control spreadsheets. The set of excerpts referring to each research question supported both the elaboration of the corresponding answers and the gathering of the strengths and weaknesses of the state-of-the-art in interactive clustering.
- E. *Quality criteria* Quality Criteria (QC) were defined (cf. Table 5) to evaluate the selected primary studies. The quality criteria evaluated the primary studies in terms of the quality of both the reporting of research results and the research per se (rigor, credibility and relevance) (Dyba et al. 2007). The quality criteria were applied considering the following possibilities: criterion *not met*, criterion *partially met* and criterion *fully met*. We used quality criteria only to characterize primary studies and not as additional exclusion criteria, since Dyba et al. (2007) argue that there is no consensus on whether or not to limit the number of primary studies in a literature review taking into account the quality of such studies.

Table 5 Quality Criteria (QC), based on Dyba et al. (2007)

Research results reporting quality	
QC-1	The study systematically presents with details: motivation, goals, research method, results, analysis, limitations and comparison with related work
Research quality	
Rigor Is the approach complete and adequate to solve the problem?	
QC-2	Decision makings and choices made in the study (e.g., choice of clustering algorithms, choice of how to include expert supervision etc.) are explicit and justified
QC-3	Results achieved are assessed using quality assessment metrics (i.e., clustering validation indices or supervised learning quality indices)
QC-4	The results achieved are compared with other clustering approaches
QC-5	The validity and significance of the results obtained are statistically attested
Credibility Is the method reproducible and verifiable, and are the results generalizable?	
QC-6	Experiments of the proposed approach are conducted with more than one dataset to avoid dependence on the results of a specific dataset
QC-7	The study is reproducible, i.e., the steps and resources used are described in detail and publicly accessible when relevant (e.g., code, raw data, summary data etc.)
Relevance Are the results useful to the scientific community and the software industry?	
QC-8	The motivations for the application of interactive clustering are explicit, i.e., the application of clustering with the inclusion of expert supervision is justified
QC-9	The quality of clustering results is discussed considering the inclusion of expert supervision, i.e., the results of interactive clustering and classic clustering are compared
QC-10	The study discusses whether the cost-benefit of including expert supervision justifies such inclusion

5.2 Conduction of the scoping review

The scoping review was conducted in three phases: (a) search for candidate studies, (b) selection of primary studies and (c) data extraction and synthesis of results. These phases are described as follows:

- a. *Search for candidate studies* To find candidate studies for primary studies, the search string was applied over the *title*, *abstract* and *keywords* fields, following the experience related by Tubío et al. (2009), in the *Scopus* search engine. The base string (cf. Table 2) was extended with additional filters that allowed automatic application of the exclusion criteria EC-1, EC-2 and EC-3 to restrict the publication year (≥ 2007), writing language (English only) as well as the fields in which the papers could be associated (computer science, engineering, mathematics, decision sciences, or business, management and

accounting only). This phase was carried out in September 2018 and resulted in 724 candidate studies.

- b. *Selection of primary studies* Applying the inclusion and exclusion criteria resulted in the selection of primary studies. For most cases, the criteria could be applied based on the analysis of the paper titles and abstracts only. When that was not enough, the full content of the paper was explored to ensure confidence in decisions about inclusion or exclusion. As a result of this phase, 50 primary studies were selected, representing 7% of the initial set of candidate studies.
- c. *Data extraction and synthesis of results* Extraction and synthesis of results, including quality analysis, were performed through a detailed analysis of the full content of the primary studies. The results of this last phase are presented in Sect. 6.

5.3 False positive examples of studies returned

We defined a very flexible search string to avoid missing relevant results. However, this string returns studies on the boundary between interactive clustering and interactive data analysis, which should be removed using the defined inclusion criteria. In addition, we used in our study a narrow definition for interactive clustering, as presented in Sect. 2.

To more accurately describe the scope of this review, Table 6 shows examples of studies returned by applying the defined search string but discarded by applying the inclusion criteria. For each example, the following are presented: (a) an excerpt from the paper abstract that allowed it to be returned as a result of the search and (b) the type of interaction discussed in the paper. These false-positive papers usually address the expert's insertion into the knowledge discovery process before (data pre-processing) or after (results post-processing) the clustering phase.

6 Results and discussion

This section presents the results of our scope review as well as a discussion of those results based on the analysis of the content extracted from the 50 primary studies selected. First, we present an overview of the interactive clustering area outlined by the reviewed studies. Then, we presented the quality evaluation of the primary studies included in this review. Finally, we answer our research questions.

6.1 Interactive clustering: an overview

The 50 primary studies selected for this scoping review (cf. Sect. 5.2) are listed in "Appendix 1" (Tables 9 and 10), organized by type of publication (journal and conference) and in descending order of the year of publication. Figure 3 shows the distribution of publications over time⁴. These data show a growth trend in the interest in interactive clustering. Although the cumulative number of publications in conferences is greater than in journals (62% to 38%), assuming a linear regression, one can infer a slightly more accentuated growth trend for publications in journals.

⁴ For 2018, only publications indexed until September were analyzed.

Table 6 False positive examples of studies returned

References	Paper abstract excerpt	Type of interaction addressed in paper
Alagambigai et al. (2008)	(...) interactive clustering in distributed environment (...)	Expert interaction occurs via a 2D data visualization to facilitate the selection of clusters in subsets of data. Expert interaction also occurs during the visualization of the final results (after the clustering phase) to facilitate data interpretation, validation and refinement
Zhang et al. (2012)	(...) we propose the interactive multiscale-nested clustering and aggregation framework to support trade space exploration (...)	Expert interaction occurs in selecting the features (dimensions) to be used in clustering, selecting the most appropriate way to view the clustering results and adding descriptive information to describe the formed clusters
Jang et al. (2014)	(...) applying interactive clustering and visualization techniques to motion tracking data (...)	Expert interaction occurs after the clustering phase through a visualization tool with the primary purpose of selecting patterns, to be saved in a gesture pattern database, which can later be used in classification tasks
Zhang et al. (2016)	(...) users to discover patterns that can be concealed by traditional global clustering via several interactive visualization techniques (...)	Focusing on manipulation and exploration of geo-localized data, expert interaction is used to select the geographic area to perform clustering, with a pre-definition of all data to be clustered. In addition, expert interaction is used for results visualization, i.e., after completion of the clustering
Behr et al. (2016)	(...) visualizing membership in latent clusters with a native interactive D3.js visualization (...)	Expert interaction occurs after the clustering phase through a visualization tool with the primary purpose of optimizing the solution to assignment problems in the context of genetic data analysis

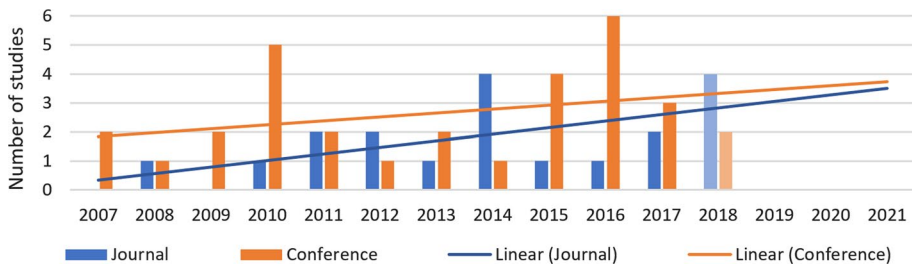


Fig. 3 Distribution over time of the primary studies

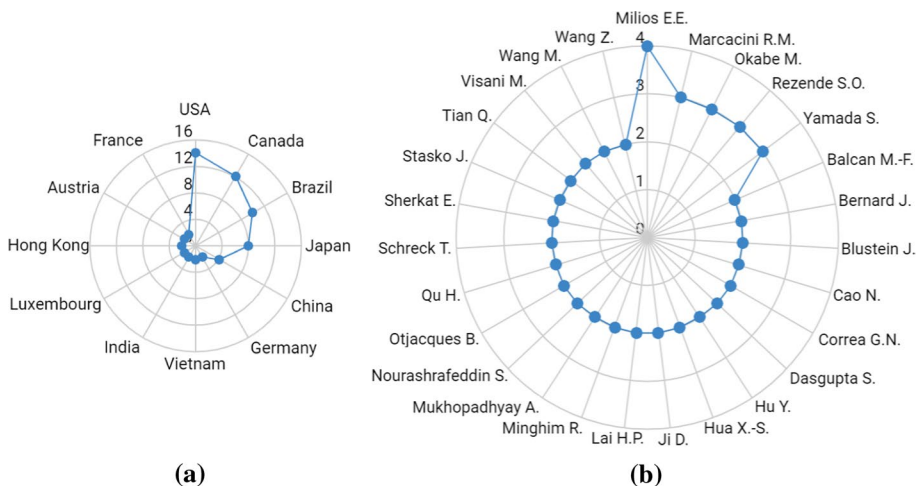


Fig. 4 Numbers of studies by authors (considering authors who have published at least two studies) and by countries

Figure 4 shows the numbers of primary studies by authors and by countries. Figure 4a shows the countries where the studies were conducted. Besides the United States of America, which is the most frequent, Canada, Brazil and Japan also stand out as the seat of part of the studies conducted in interactive clustering. Figure 4b shows the authors who have published the most in interactive clustering, considering only those who published at least two studies. Taking only the five most productive authors, the following two main interest groups in interactive clustering stand out:

- Researchers E. E. Milios (Dalhousie University, Canada), R. M. Marcacini, S. O. Rezende (University of São Paulo, Brazil), and their collaborators, are especially interested in applying interactive clustering to textual data. The inserted expert's knowledge is used to: select features in an iterative process of refining the clusters resulting from clustering (Hu et al. 2011, 2014); and apply such an iterative feature selection process to cluster the terms (features) that describe documents (data points). Discriminative information extracted from term clustering is used to support initialization procedures in the document clustering process (Nourashrafeddin et al.

Fig. 5 Word cloud for topics covered in interactive clustering research



2018; Sherkat et al. 2018); and rank features to implement feature enrichment procedures (Marcacini et al. 2012, 2013; Correa et al. 2015).

- Researchers M. Okabe (Japan University of Technology) and S. Yamada (National Institute of Informatics, Japan) address structured data on agriculture, floriculture, and web pages (Okabe and Yamada 2010a, b, 2011). They focus on the representation of the knowledge provided by the human experts through must-link/cannot-link constraints. Using these types of constraints, they develop algorithms that allow the learning of similarity relationships between data, while considering such constraints.

Figure 5 shows, through a word cloud, an analysis of the keywords associated by the authors to their papers. Words referring to the same context were grouped (e.g., *explainability* and *interpretability* as well as *image analysis* and *image indexing*). The size of each word in the cloud refers to the frequency with which the corresponding concept appears in the keyword list of primary studies.

Analysis of the word cloud reveals the great attention given to *visualization*. The interaction with human experts requires a mechanism that allows them to visualize the clustering results. Thus, visualization stands out as an important requirement for the development of the interactive clustering area. One can also notice through the word cloud a large occurrence of keywords related to applying interactive clustering specifically to the *text and image data* domains. Another topic that draws attention in the word cloud is the *feature selection task*, revealing that the influence of the expert's knowledge in the organization of the data vector space is one of the main topics in interactive cluster. Finally, there is also an emphasis on keywords related to *explainability*, showing that interactive clustering can contribute to the current general interest in explainable artificial intelligence.

6.2 Quality evaluation of primary studies

Table 7 shows the result of the quality evaluation of primary studies, carried out by applying the quality criteria and the evaluation system set out in Sect. 5.1. Figure 6 shows the distribution of evaluations in relation to each quality criterion.

The vast majority of primary studies fully met the criterion related to the overall quality of the report (QC-1): 40 out of 50 (i.e., 80%). Studies that only partially met this criterion presented one or more problems related to: (i) lack of clarity in defining the goals of the study, the applied research method, or the research limitations; (ii) lack of

Table 7 Quality evaluation: fully met (+); partially met (-); not met at all (none)

Primary study	Quality criteria (QC)									
	Rep.	Rigor				Credib.		Relevance		
	1	2	3	4	5	6	7	8	9	10
Arin et al. (2018)	+	+	+	+	+		+	+	+	
Cavallo and Demiralp (2018)	+	+	+	+		+	+	+	+	
Mai et al. (2018)	+	-	+	+		+	+	+	+	-
Nourashrafeddin et al. (2018)	+	+	+	+		+	+	+	+	+
Sacha et al. (2018)	-	+				+	-	+		
Sherkat et al. (2018)	+	+	+	+		+	-	+	+	-
Awasthi et al. (2017)	+	-	+	+		+	+	+		
Coden et al. (2017)	+	-	+	+		+	-	+		+
Emamjomeh-Zadeh and Kempe (2017)*	-	-					-	+		
Ferrero et al. (2017)	-							+		
Lei et al. (2017)	-	+	+	+		+	+	+	+	
Boudjeloud-Assala et al. (2016)	+	+	+	+	+	+	+	+	+	
Chang et al. (2016)	-	-	+					+	+	
Mauder et al. (2016)	-	-				+	-	+		
Mukhopadhyay (2016)	+	+	+			+		+		
Vikram and Dasgupta (2016)	+	+	+	+		+	+	+		
Vu et al. (2016)	+	+	+	+			+	+		
Xu et al. (2016)	+	-		+		+		+	+	
Correa et al. (2015)	+	-	+	+	+	+	+	+	+	+
Gieseke et al. (2015)	+	+		+		+	+	+	+	
Khodabandeh et al. (2015)	+	+	+	+		+	+	+	+	
Lelkes and Reyzin (2015)*	+	+					+	+		-
Senderovich and Maysuradze (2015)	+	+	+	+		+		+		+
Geerts and Ndindi (2014)	+	-		+		+	-			
Hu et al. (2014)	+		+	+		+	+	+		+
Lai et al. (2014)	+	+	+	+	+	+	+	+	+	
Xiong et al. (2014)	+	+	+	+	+	+	+	+		+
Zhang et al. (2014)	+	-	+	+			+	+	+	
Bruneau and Otjacques (2013)	+	+	+	+	+	+	-	+		
Marcacini et al. (2013)	-	-	+			+	-	+	+	
Mukhopadhyay et al. (2013)	+	-	+	+	+	+	-	+	+	
Lee et al. (2012)	+	+		+			-	+	+	
Marcacini et al. (2012)	+	-	+	+		+	+	+	+	+
Wang et al. (2012)	+	+	+	+		+	-	+	+	
Cao et al. (2011)	+	+			+	+	-	+		
Fredj et al. (2011)	-					+		+		
Hu et al. (2011)	+	-	+	+		+	+	+	+	+
Okabe and Yamada (2011)	+	+	+	+		+	-	+		+
Dasgupta and Ng (2010)	+	-	+	+		+	+	+		
Dubey et al. (2010)	+	-	+	+		+	+	+	+	
Guo et al. (2010)	+	+		+		+		+		-
Ji et al. (2010)	+			+			-	+	+	-

Table 7 (continued)

Primary study	Quality criteria (QC)									
	Rep.	Rigor				Credib.		Relevance		
	1	2	3	4	5	6	7	8	9	10
Okabe and Yamada (2010a)	–	–	+	+		+	+	–		
Okabe and Yamada (2010b)	+	–	+	+		+		+		
Momma et al. (2009)	+	–	+	+		+	+	+		
Schreck et al. (2009)	+	+		+				+		
Balkan and Blum (2008)*	+	+					+	+		
Zhu et al. (2008)	+	+	+	+			+	+	+	
desJardins et al. (2007)	+	–	+	+		+	+	+		
Iorio et al. (2007)	–		+	+		+	–			

*These studies formalize interactive clustering strategies and define classes of problems to assess the efficiency of the corresponding strategies. Although they do not satisfy QC-6 as proposed in our review protocol, they have a good level of generalization considering applications that can be modeled in such classes of problems

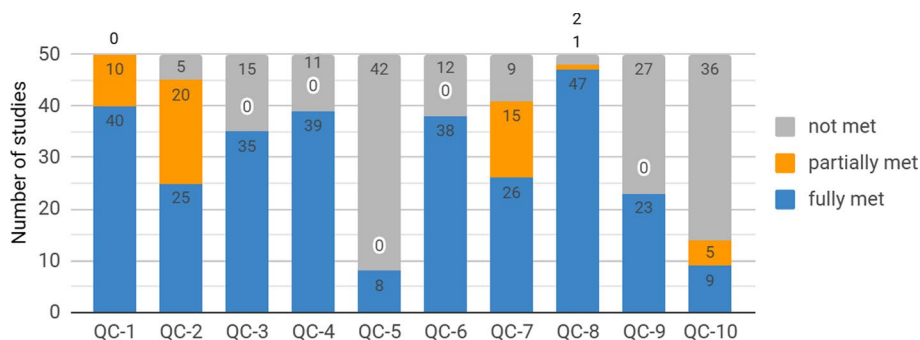


Fig. 6 Distribution of quality criteria evaluations. The numbers associated with the colored bands (in each bar) show the number of primary studies that fully met the corresponding criterion (in blue), partially met the corresponding criterion (in orange) and not met the corresponding criterion at all (in gray). (Color figure online)

systematization of information, which makes it difficult for the reader to understand the proposed approaches and research strategies followed; and (iii) superficiality of the information presented, although often justified by space restrictions. The evaluation according to this criterion shows that these works are, in general, good research reports, whose reading provides an adequate understanding of the study carried out in interactive clustering.

Scientific rigor was met by most studies for criteria QC-3 and QC-4, related to respectively *assessment of results with quality metrics* and *comparison with other clustering approaches*. However, QC-2 was fully met by only 50% of the studies and QC-5 was not met by the vast majority of studies, not even partially (42 out of 50, i.e., 84%). Partial or non-compliance with QC-2 results from the limited description of the applied research method (as also pointed out in the evaluation of QC-1). The main flaws in meeting this criterion is the lack of justification for choosing the algorithms and datasets used in the study.

The low level of compliance with QC-5 shows a worrying low statistical credibility of the results presented in the interactive clustering studies.

The credibility criterion QC-6 was fully met by 38 studies (76%), showing a frequent concern of the authors in achieving some degree of generalization for their proposed approaches to interactive clustering. On the other hand, the credibility criterion QC-7, related to reproducibility⁵, was fully met by only 52% of the studies, showing a weakness in interactive clustering research.

Finally, the evaluation of the relevance criteria showed a contradiction between what is proposed and what is delivered by the results and discussions presented in the scientific papers in this area. The vast majority of studies fully met QC-8 (94%), i.e., the authors successfully motivated the inclusion of supervision in the clustering process. However, according to the evaluation of QC-9, only 23 studies (46%) compared the results obtained through the expert supervision with results obtained without this supervision. In addition, only 14 studies (28%) met at least partially QC-10, which addresses the analysis of the cost-benefit ratio of the insertion of expert supervision in the clustering process. In five studies that only partially met this criterion, the authors only expressed some concern about decreasing the expert's effort during the supervision task.

This quality assessment shows that the proposed approaches are overall well defined and properly motivated (cf. QC-1 and QC-8) and have promising results (cf. QC-4 and QC-6). However, on the other hand, the way the results are assessed (issues in QC-5, QC-9 and QC-10) are still insufficient to adequately define the effectiveness of interactive clustering.

6.3 Research questions discussion

Below are answers to each of the five research questions. Answers are discussed based on graphs that summarize information extracted from primary studies. Tables with detailed data are presented in "Appendix 1".

6.3.1 Data types and application domains (RQ-1)

To answer this research question, four types of information were extracted from each primary study: (a) the data types used in the reported experiments; (b) whether the interactive clustering approach presented is independent of the data types used in the experiments; (c) whether the interactive clustering approach presented is independent of the context (process) that generated the data used in the experiments; and (d) the application domain for which interactive clustering was modeled. The data extracted to answer this research question are detailed in Table 11 of "Appendix 1".

An interactive clustering approach was classified as *data type independent* in two situations: (a) if the approach was experimented on datasets of different data types; or (b) if the authors present explicit statements about the possibility of using their approach on data types other than those used in their experiments. Likewise, an approach was classified as *context independent* if it was experimented with more than one dataset, even if of the same data type. In this case, the diversity in the datasets would prevent statistical characteristics specific of a single dataset from being the reason for the clustering results obtained.

⁵ The reproducibility evaluation disregarded the need for a sample of experts statistically equivalent to that of the study carried out.

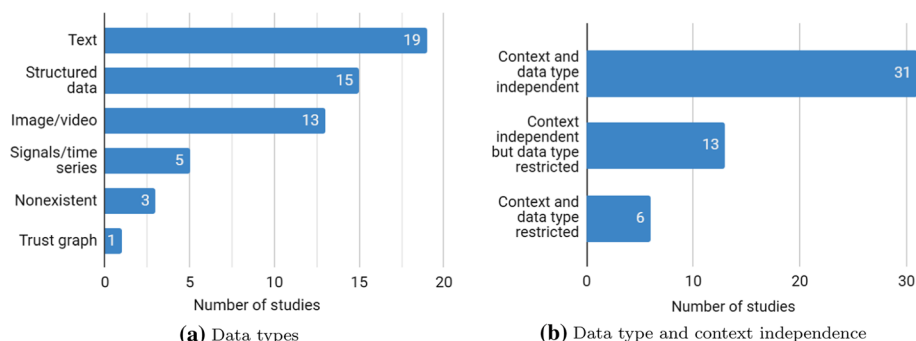


Fig. 7 RQ1—data types and independence on data type and context

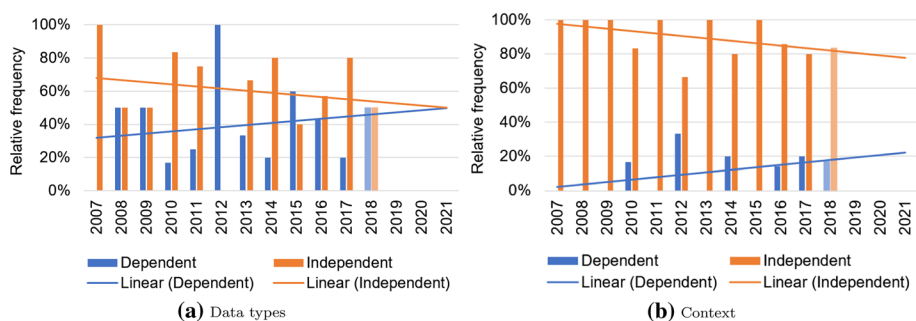


Fig. 8 RQ1—dependency relationships for data types and context

Figure 7a shows a summary of the data types used in the experiments reported in the primary studies⁶. According to the taxonomy used for data types, *structured data* refers to those that we could assume be represented by a collection of independent and identically distributed⁷ random variables (i.e., features). The analysis of the figure shows that textual data are the most often used in experiments (19 studies, 38%), followed by structured data (15 studies, 30%) and image/video (13 studies, 26%). The results of this analysis of data types corroborates the analysis of research topics derived from the keywords used by the authors in their papers. In the word cloud (cf. Fig. 5), *text-mining* and *image-analysis* are respectively the fourth and sixth most cited topics within 32 topics considered.

The reason why these data types are the most often used could not be extracted from the studies; however, we raised the following hypotheses for textual data and images/videos: (a) the analysis of textual data is closely linked to natural language processing and semantic interpretations, which still have many challenges and depend on human work for tasks such as labeling, interpretation and validation of automated analysis results; and (b) the analysis of images/videos is conducive to involving human experts, mainly because it is a type of information easy to be represented and analyzed visually, motivating and facilitating the proposition and testing of approaches to insert human collaboration.

⁶ As there are studies that present experiments with different data types, the sum accumulated in the graph is higher than the number of studies analyzed.

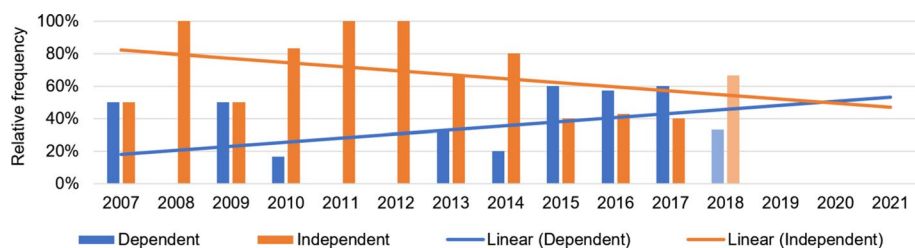
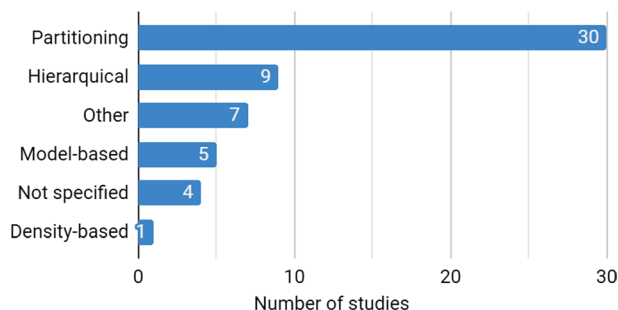
⁷ independent and identically distributed (i.i.d.)—from the probability theory and statistics.

Table 8 RQ1—application domains

Application domains	Number of studies
News article topic analysis	13
Medical diagnosis support	7
Botanical data classification, paper topic analysis	6
Agricultural decision support, handwritten digits recognition, web pages topic analysis	4
Biological data classification	3
Criminal analysis support, object detection in images, photo album organization, retail decision support, satellite image analysis, sentiment analysis	2
Audio analysis, bio-archaeological data classification, digit recognition, face recognition, financial decision support, hand movement classification, healthcare decision support, human activity recognition, scientists' publishing behaviour analysis, sign language processing, question and answer forum analysis, remotely sensed data analysis, search result topic analysis, signal processing, social network interaction analysis, business decision support, survey response analysis, legislators' voting behaviour, zoo-technical classification	1

Figure 7b shows a summary of information on data type and context independence. We consider data context as characteristics that further specialize a data type. For example, image clustering can address tasks in images in general or tasks in specific images such as with people's faces; the latter known as face clustering. Thus, face clustering depends on both the data type (images) and the data context (images with people's faces). Another example of context dependency is tweet clustering in contrast to clustering of text in general. The high levels of data type independence (31 studies, 62%) and context independence (44 studies, 88%) show that the authors have shown concern in evaluating the generalization of their approaches. However, an over time analysis of these data (cf. Fig. 8a, b, respectively) shows that the levels of independence on data types and independence on context have been slightly lower in recent years. In fact, the trend lines associated with independent in both graphs show a downward trend, representing a point of attention for researchers in this area.

Finally, Table 8 lists the application domains related to the data used in the experiments reported in the primary studies. There are a wide variety of application domains (33 in total). This variety may be due to the large use of public domain benchmark datasets, which are collected from multiple domains. For example, the *botanical data classification* domain refers to the Iris dataset, one of the datasets most often used for studies and examples of data analysis strategies in general. The authors of the studies analyzed very rarely justify the choice of the application domains, which may show that they are not really interested in any particular domain, but only in testing or exemplifying their approaches. Thus, it would not be reasonable to conclude that all of these domains used are susceptible to the application of interactive clustering strategies. Despite this, three of the most often used domains (news articles topic analysis, papers topic analysis and medical diagnostic support) are in fact directly related to data types of interest in the interactive clustering context, as showed in Figs. 5 and 7.

Fig. 9 RQ2—clustering algorithm classes**Fig. 10** RQ2—dependency relationship between interactive clustering approaches and clustering algorithms

6.3.2 Clustering algorithm classes (RQ-2)

Figure 9 shows a summary of the clustering algorithm classes found in the primary studies. The most often used algorithm classes are *partitioning* algorithms (30 studies, 60%), represented mainly by k-means, and *hierarchical* algorithms (9 studies, 18%), represented all by bottom-up hierarchical clustering strategies. *Model-based* algorithms (expectation-maximization and self-organizing maps) and based on *other* approaches (e.g., graph theory and genetic algorithms) are also present, albeit at a lower frequency. *Density-based* algorithms were cited only once, in a study that applied algorithms from several classes of clustering algorithms. The data extracted to answer this research question are detailed in Table 12 of "Appendix 1".

Of the 30 studies using partitioning algorithms for interactive clustering, only two motivated the choice of this class of clustering algorithms, which were done based on its simplicity, efficiency and effectiveness and its consequent popularity for classic clustering; while the remaining 28 studies present no motivation. Similarly, of the nine studies using hierarchical algorithms, only three motivated the choice of this class of algorithms, which were done based on: (i) ease of creating clustering result visualizations using the hierarchy of clusters produced by the algorithm; (ii) no need to define the number of clusters in advance; and (iii) ease of applying split/merge requests and the algorithm's efficiency and effectiveness.

The ease of creating visualizations for the resulting clusters as offered by hierarchical algorithms addresses the key concern in the general context of interactive clustering: *visualization* (see Fig. 5). In addition, easing the application of split/merge requests is relevant as it is one of the most often used strategies for inserting expert's knowledge into the clustering process. The algorithm efficiency is also a relevant aspect due to the need to insert human experts in the clustering process loop. The clustering algorithm

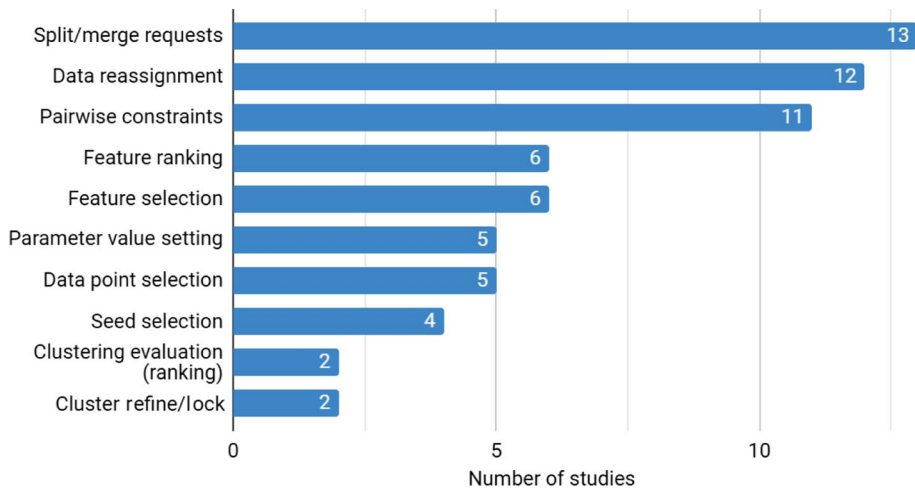


Fig. 11 RQ3—strategies for acquiring expert's knowledge

needs to be efficient so that the intermediate results of the clustering process are delivered quickly so as not to render the expert idle. On the other hand, the other justifications found (i.e., simplicity, classic clustering-related effectiveness, popularity and no need to define parameters) refer to the general context of classic clustering and hence do not directly address the specifics of interactive clustering.

Figure 10 shows the dependency relationship between interactive clustering approaches (i.e., strategies for interacting with experts and inserting their feedback in the clustering process) and clustering algorithms. Most primary studies (33 studies, 66%) propose interactive clustering approaches to work independently of the chosen clustering algorithm. However, interactive clustering approaches created especially for use with a particular clustering algorithm have become more frequent in recent years, as shown in the trend line in Fig. 10. This trend highlights that more specific solutions have been explored, with less application flexibility, as also observed in the analysis of Fig. 8a, b.

6.3.3 Expert supervision (RQ-3)

To answer this research question, the following aspects were analyzed: (a) how the expert's knowledge is acquired; (b) how the acquired expert's knowledge is inserted into the clustering process; (c) which strategies are applied for presenting elements for analysis to the expert; and (d) whether the interaction is synchronous or asynchronous. The data extracted to answer the first three aspects are detailed in Tables 13, 14 and 15 of "Appendix 1". For the fourth aspect, no accompanying table with the extracted data is presented as virtually no primary study explicitly presents the form of interaction used.

A. How expert's knowledge is acquired

Figure 11 shows the classification regarding the types of strategies in which knowledge is extracted from interaction with the expert^{8,9}.

Split/merge requests is the strategy most often found in primary studies (13 studies, 26%). Of these 13 studies, two use only merge requests due to characteristics inherent to the problem addressed and the clustering solutions applied. Split/merge requests are requests to split one cluster into two or more clusters or merge two or more clusters into one cluster. Experts may request to split a cluster because they realize that the subset of data points associated with it seems to characterize more than one behavior profile in the application domain. On the other hand, experts may ask to merge clusters by the reverse reasoning, i.e., different clusters are formed by subsets of data points that seem to characterize a same behavior profile in the application domain.

The second most often found strategy for knowledge acquisition is *data reassignment*, in which the expert changes the cluster to which a data was initially associated by the clustering algorithm. *Pairwise constraints* also occur with a high frequency in the primary studies. Through this strategy, expert's knowledge is acquired in the form of constraints imposed on pairs of data points, showing whether they must be necessarily associated with the same cluster (i.e., *must-link constraints*) or whether they must be necessarily associated with different clusters (i.e., *cannot-link constraints*).

Although split/merge requests, data reassignment and pairwise constraints are common strategies in interactive clustering, the authors report that they are costly for the expert and the success of its application depends on how the information about the clusters is made available to the expert, in particular for the last two strategies. Ideally, all data and their allocations to the clusters should be viewed together.

Feature selection is a knowledge acquisition strategy in which the expert chooses the most relevant features to describe the data to be clustered. Similarly, through *feature ranking*, the expert ranks the features in terms of how important they are to describe the data to be clustered. Both strategies were found in six studies (12%) each.

Parameter value setting (5 studies, 10%) and *seed selection* (4 studies, 8%) are strategies for knowledge acquisition that require a certain level of technical knowledge on clustering techniques by the expert. For parameter value setting, the expert is asked to adjust the parameters used in the algorithms (e.g., definition of the similarity threshold used in hierarchical clustering algorithms). As for seed selection, the expert must select data points to support the initialization of the clustering algorithms, which occurs mainly with partitioning algorithms in which the initialization is done through the determination of prototype vectors.

Data point selection (5 studies, 10%) refers to the choice by the expert of data points from the dataset for submission to a separate clustering, specifically for those data points. *Clustering evaluation (ranking)* (2 studies, 4%) requires the expert to evaluate and rank the suitability of different solutions, i.e., the results of different clusterings. Finally, *cluster refine/lock* (2 studies, 4%) refers to the expert to inform whether: a cluster has not yet reached an adequate quality level in relation to the data points that comprise it and hence must still be submitted to the clustering algorithm (refine); or a cluster has reached the quality level expected and hence should not undergo further modifications (lock).

⁸ Only acquisition strategies found in at least two primary studies are shown in the figure.

⁹ As there are studies that present experiments with different strategies for acquiring knowledge, the sum accumulated in the graph is higher than the number of studies analyzed.

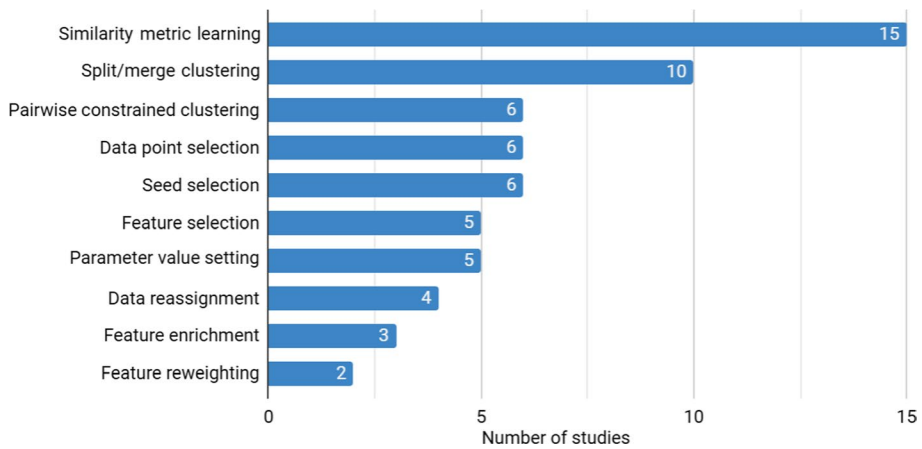


Fig. 12 RQ3—strategies for inserting the expert's knowledge

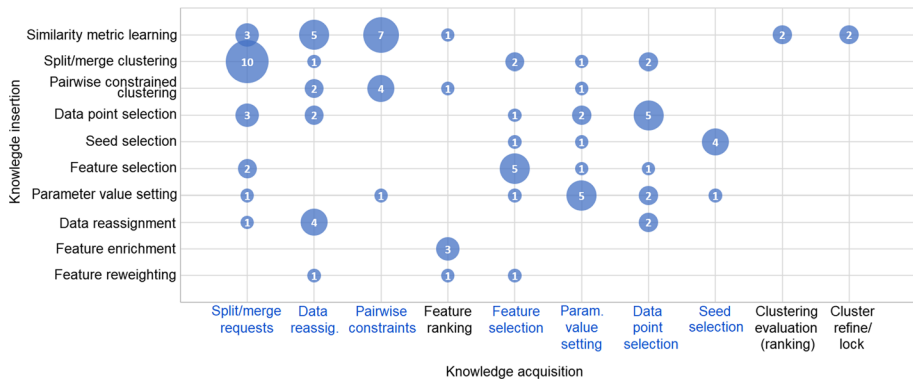


Fig. 13 RQ3—combinations between strategies for acquiring the expert's knowledge and strategies for inserting that knowledge into the clustering process

B. How acquired expert's knowledge is used

Figure 12 shows the classification regarding the types of strategies in which the knowledge acquired from the expert is inserted to be used into the clustering process^{10,11}.

Similarity metric learning is the strategy for inserting the expert's knowledge most often found in primary studies (15 studies, 30%). Through this strategy, the expert's knowledge of the similarity between data points is used to model a distance function to quantify the similarity relationships between data points. As the implementation of similarity metric learning does not depend on the strategy used for acquiring expert's knowledge, then this strategy for inserting the expert's knowledge is applied in conjunction with various

¹⁰ Only strategies found in at least two primary studies are shown in the figure.

¹¹ As there are studies that present experiments with different strategies for using the acquired knowledge, the sum accumulated in the graph is higher than the number of studies analyzed.

strategies for acquiring expert's knowledge, as shown in Fig. 13. This strategy replaces arbitrary choices of a similarity measure to be used in the clustering algorithm and is especially useful when the expert's knowledge is acquired using pairwise constraints or data reassignment. When using these knowledge acquisition strategies, the expert commonly indicates associations between data points that the clustering algorithm would not indicate following an arbitrarily chosen distance measure.

Split/merge clustering (10 studies, 20%) and *pairwise constrained clustering* (6 studies, 12%) are the following most often found strategies, possibly due to their high degree of maturity. These two strategies for inserting expert's knowledge into the clustering process gave rise to two strategies for knowledge acquisition, respectively, *split/merge requests* and *pairwise constraints* (cf. Fig. 11). Despite this association, the knowledge acquired through those strategies (split/merge requests and pairwise constraints) does not necessarily need to be inserted into the clustering process through, respectively, these techniques (split/merge clustering and pairwise constrained clustering), as shown in Fig. 13. Some authors point out that pairwise constrained clustering is effective only when very well modeled, as inappropriate constraints can lead to worsening clustering results. In addition, authors also warn of the high cost of creating and applying constraints and the risk of cluster over-merging when using pairwise constrained clustering.

Data point selection (6 studies, 12%), *seed selection* (6 studies, 12%), *feature selection* (5 studies, 10%), *parameter value setting* (5 studies, 10%) and *data reassignment* (4 studies, 8%) are strategies for inserting expert's knowledge into the clustering process directly related to equivalent strategies for expert's knowledge acquisition (cf. Fig. 11). Thus, each of these five insertion strategies corresponds to the direct application of the expert's knowledge in the interaction loop with the clustering algorithm, without changing the logic of the algorithm.

Feature enrichment (3 studies, 6%) and *feature reweighting* (2 studies, 4%) are used to, respectively, build new features or weight features using the expert's knowledge acquired mainly through the *feature selection* or *feature ranking* strategies (cf. Fig. 11).

Figure 13 shows the combinations between strategies for acquiring the expert's knowledge and strategies for inserting that knowledge into the clustering process. The size of each circle as well as the number within it represents the number of primary studies that have that respective combination of acquisition and insertion strategies. The observed distribution of circles in the graph shows that research on interactive clustering is still in the exploratory stage, as there seems to be no combination patterns representing well-established general strategies that could be considered best practices in this area. However, from the current picture, we can highlight that:

- The most frequent combinations are between *split/merge requests* and *split/merge clustering* and between *pairwise constraints* and *similarity metric learning*. Although these are the most frequent combinations, these numbers (10 and 7) neither show a very representative pattern nor a high level of maturity, mainly when considering the relatively small sample of primary studies and the high number of possible combinations between strategies. Despite this, these two combinations can show starting points for inexperienced researchers in the interactive clustering area.
- Although strategies for inserting knowledge and strategies for acquiring knowledge could be expected to appear exclusively combined when of the same nature (e.g., seed selection), this did not occur as a general rule. Considering the seven strategies for knowledge acquisition (with *blue* labels on the graph) that have strategies directly associated with knowledge insertion, there are 86 combinations of them with any knowl-

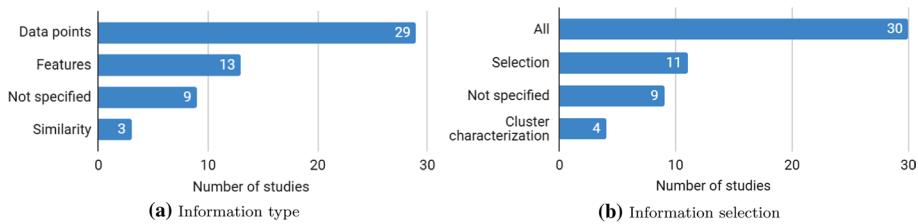


Fig. 14 RQ3—strategies for presenting information for supervision

edge insertion strategies. However, of these 86 combinations, only 37 (43%) refer to combinations of them with their homonym counterparts in the knowledge insertion strategies.

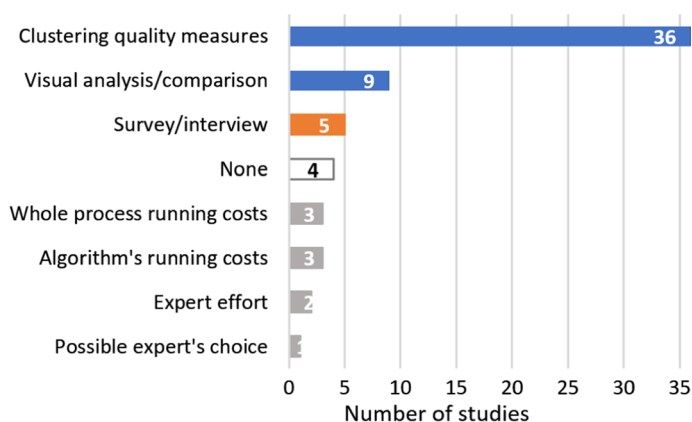
- Knowledge acquisition strategies *data reassignment* and *pairwise constraints* shows a combination pattern different from the other ones. These two acquisition strategies have more combinations with *similarity metric learning* than with their homonym counterparts.

C. Strategies for presenting clustering information

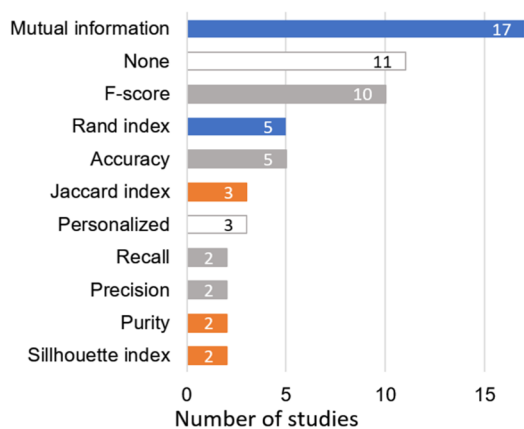
Figure 14a shows the frequencies of the strategies used to present information about the clustering result to the expert. The strategies were classified into the use of: *data points* being clustered, in their original form or using abstractions or projections (29 studies, 58%); *features* characterizing the data points, through their labels/meanings or through abstractions or projections (13 studies, 26%); and information about *similarity* between the data points, through its simplest form (i.e., similarity matrix) or through abstractions (e.g., dendograms) or projections (e.g., U-matrix, which is a projection of similarity and neighborhood relation between units of a self-organizing map (Kohonen 2000; Ultsch and Siemon 1990)) (3 studies, 6%). Of the 50 primary studies, nine studies (18%) did not report how the clustering results were presented to experts. The development of graphical interface tools dedicated to presenting information to the expert was mentioned in 26 studies (52%).

When presenting information to the expert, the data analyst can choose between presenting all information related to the clustering under analysis or selecting only a part of that information. Selecting only a part of the information can be useful to optimize the supervisory work in three aspects: (i) decreasing the time spent analyzing the information; (ii) making work more effective by presenting information with a greater chance of positively impacting the quality of results; or (iii) improving the performance of the clustering process. As shown in Fig. 14b, *all* information was presented in 30 studies (60%) and part of the information was *selected* for presentation in 11 studies (22%). In addition, in four studies (8%), part of the information was selected, but with the clear intention of the researchers to select the necessary and sufficient information to *characterize each of the resulting clusters*.

When *image* data is being analyzed in the cluster, the visualization is done using the data itself. For other data types, visualization depends on abstractions built with color coding or projections in two or three dimensions using *multidimensional scaling* (Borg and Groenen 1997), *principal component analysis* (Wold et al. 1987) and *t-SNE* (van der



(a) Expert supervision evaluation strategies



(b) Clustering validation measures

Fig. 15 RQ4—evaluation strategies

Maaten and Hinton 2008). Of the 50 primary studies, 16 studies (32%) report the use of abstractions or projections for data visualization. From the expert's point of view, analyzing the data itself (as for images) may be more natural than analyzing abstractions.

D. Strategies for interacting with human experts

It is difficult to accurately extract, from the reading the primary studies, whether the expert's interventions in the clustering process occurred synchronously (i.e., online) or asynchronously (i.e., offline). Some of the studies imply that interactions with the expert occurred in a synchronous way, i.e., during each supervisory iteration, the expert would be able to visualize in a timely manner the results produced by the clustering algorithm. This possibility is somewhat more evident in the 26 studies reporting the development of a graphical interface tool to support the interactive clustering process.

Synchronous interaction with the expert can be feasible when the waiting time for the presentation of new clustering results is short enough to not leave the expert idle in the clustering process. Thus, knowledge about the asymptotic complexity of clustering algorithms is important when proposing interactive clustering approaches and deciding to use synchronous or asynchronous supervision. High asymptotic complexity would make synchronous interaction unfeasible. However, this issue is discussed in only eight studies (16%), with only one of them presenting an interaction supported by a graphic tool to visualize the clustering results.

6.3.4 Evaluation of expert supervision and clustering outcomes (RQ-4)

To answer this research question, an analysis was conducted on how the authors evaluated the expert's supervisory work against the quality of the obtained clustering results. The data extracted to answer this research question are detailed in Tables 16 and 17 of "Appendix 1".

Figure 15a shows an overview of the evaluation strategies used in primary studies¹². This evaluation is overall made from three different perspectives: *quality of clustering results* (in blue), *meeting the expert's expectations* (in orange) and *the supervision effect on the clustering process* (in gray).

According to the data in Figure 15a, most studies (36 studies, 72%) apply *clustering quality measures*. In 23 studies (46%), only clustering quality measures are used as an evaluation strategy. The prevalence of this type of evaluation shows that studies being conducted in interactive clustering have not focused on evaluating whether the expert's expectations on the quality of the results are met and whether the expert's participation is being carried out effectively.

The second most frequent evaluation strategy, with 9 studies (18%), is *visual analysis and comparisons*, used to inspect the visualization of the clustering results alone or in comparison with the results obtained in two consecutive process iterations. *Surveys and interviews* (9 studies, 18%) are strategies to ask the expert what is their opinion and observations on the results obtained.

The last group of evaluation strategies seeks to evaluate whether: the expert's supervision increased the cost of executing the clustering process, i.e., many interactions with the expert were required or much information from the expert was demanded (*whole process running costs*—3 studies, 6%); the expert's supervision increased the cost of executing the clustering algorithm (*algorithm's running costs*—3 studies, 6%); all the effort expended by the expert was necessary (*expert effort*—2 studies, 4%); and different types of information provided by the expert influence clustering results differently (*possible expert's choice*—1 study, 2%).

Figure 15b shows an overview of the validation measures used in primary studies^{13,14}. These measures include: *external clustering validation indexes* (in blue), *internal clustering validation indexes* (in orange) and *measures commonly used in supervised learning outcomes* (in gray).

Several authors point out the achievement of results that are more adherent to the expectations of experts (or users) as one of the reasons for including supervision in the clustering process. However, this is an essentially subjective evaluation that cannot commonly

¹² As there are studies that present experiments using different evaluation strategies, the sum accumulated in the graph is higher than the number of studies analyzed.

¹³ Only validation measures found in at least two primary studies are shown in the figure.

¹⁴ As there are studies that present experiments using different validation measures, the sum accumulated in the graph is higher than the number of studies analyzed.

be performed by internal clustering validation indexes. These internal validation indexes evaluate the adequacy of the association between data and clusters based on the similarity measures used in the clustering algorithm as well as the organization of data points in the vector space established by the features used to describe them. If the expert's knowledge was inserted in the clustering process in the form, for example, of constraints (must-link/cannot-link) or data reassignment, the final clustering result will not necessarily correspond to that expected by these indexes. This phenomenon may justify the low occurrence of internal validation indexes in interactive clustering studies.

In addition to these evaluation strategies, 23 studies (46%) also show the concern to compare the clustering results obtained with the interactive clustering process against the results obtained with the classic clustering process. Among the 27 remaining studies, the authors of 23 studies compare with other interactive or semi-supervised approaches. One can notice that researchers in interactive clustering seem to be more motivated to seek the best strategy to use the expert's knowledge than to discuss whether the expert's supervision should really be included in the clustering process.

6.3.5 Expert's effort (RQ-5)

Interactive clustering researchers have used simulation systems to replace the human experts. Simulators can be built when labeled datasets are available for experiment running, since the labels associated with each data point can represent the expert's knowledge. Simulators were used in 30 studies, i.e., in 60% of primary studies. The use of simulators was justified by one of the authors as being a way to improve the reproducibility of the experiments. However, we can raise the hypothesis that the use of simulators may also be motivated by the cost or the unavailability of experts.

The strategies for measuring the expert's effort identified in the analysis of the primary studies are:

- Number of actions necessary to achieve the goal desired by the expert.
- Number of queries answered by the expert.
- Number of features chosen by the expert in relation to the number of features presented to them (for the case of feature selection).
- Supervision task duration.

In order to outline a discussion on the state of the research in interactive clustering with regards to this issue, we consider two scenarios:

- Of the 30 studies that replace the human expert with a simulator, 11 studies (36%) present an analysis of the expert's effort, in this case, a simulated expert.
- Of the 20 studies that rely on human experts, four studies (20%) present an analysis of the expert's effort, in this case, a human expert.

From these data, we concluded that few researchers are working on measuring the expert's effort. Considering this entire analysis scenario, only 15 studies (36%) present any analysis related to this issue. In addition, no study has reported any type of analysis on the cost-benefit ratio of including expert supervision in the clustering process.

The data extracted to answer this research question are detailed in Table 18 of "Appendix 1".

6.4 Analysis of research questions

This section presents an analytical summary of the responses produced for each of the research questions, highlighting open questions and research opportunities in interactive clustering.

- RQ-1** In what data types and application domains has interactive clustering been applied to? We identified five data types and 33 application domains reported in the primary studies, with prevalence of text and image data types and application domains related to these data types (i.e., text mining and image analysis). The use of structured data is also frequent but due to the ease of finding benchmark datasets and not necessarily to the motivation to solve given problems in particular. In addition, although 62% of studies report interactive clustering approaches independent of the data type, an analysis over time showed a trend to move towards domain-driven solutions. Finally, we observed that the data types submitted to interactive clustering approaches do not include data from process automation. These data are found in the form of event log, have very peculiar characteristics and have been studied in the area of process mining. We envision an opportunity for using interactive clustering mainly in the automation of business processes. In this application domain, the expectations of business analysts (i.e., the experts) about the results of data analysis usually differ greatly from what classical data analysis approaches can offer. The association of the business analyst with the data analysis loop and, in particular, with the clustering analysis loop represents an opportunity for development for the interactive clustering area, with a contribution to the process mining area.
- RQ-2** Which classes of clustering algorithms have been used in interactive clustering? And why were they chosen? Supported by the taxonomy of clustering algorithm classes proposed by Han et al. (2011), we organized the different clustering algorithms used in interactive clustering into four classes. However, we found no correspondence with the definitions that support this taxonomy for seven of the algorithms identified in the primary studies. Despite this, two classes of algorithms have been identified as most suitable for use in interactive clustering: partitioning algorithms, due to their effectiveness in proposing good solutions, and hierarchical algorithms, due to their effectiveness in proposing solutions in hierarchical form, which makes it easy to use split/merge strategies and to construct visualizations with high quality information about the solutions. Partitioning algorithms were found much more frequently than hierarchical algorithms, whose reason could not be understood by reading the content analyzed in this review. Thus, the reason for this supremacy of partitioning algorithms can be found by studying the relationship between the choices made by interactive clustering researchers and the availability of algorithms that allow the implementation of rules and constraints in the clustering process (e.g., constrained clustering). However, this analysis is beyond the scope of the goals of this scope review, although it represents an opportunity for future work.
- RQ-3** How has expert supervision been implemented in interactive clustering? To discuss the insertion of the expert in the loop of the interactive clustering, we proposed to categorize the information extracted from the primary studies. From the analysis of the results of these categorizations, we found that exploring the expert's knowledge to directly influence the structure of the clusters (through split/merge requests, data reassignment, pairwise constraints and similarity metric learning,

split/merge clustering and pairwise constrained clustering) is the most frequent approach in the area. A cross analysis revealed that knowledge acquisition and knowledge insertion strategies belonging to similar categories (e.g., split/merge requests and split/merge clustering) are not necessarily used together to implement a complete interactive clustering process. Another relevant aspect about the interaction with the specialist refers to how information about clustering results can be presented for the specialist's supervisory work. The categorization we proposed for this analysis showed that, in most studies, all data points in the dataset under analysis by clustering are presented to the specialist. This type of strategy burdens the work of the human specialist; however, the search for alternatives has received little attention. This is, therefore, an open question to be explored by researchers in interactive clustering aiming at offering scalability for the application of interactive clustering in domains with a large volume of data to be processed. One way to explore this issue is to propose efficient tools for visualizing cluster results, through abstractions and graphical projections. One can notice that only half of the primary studies mentioned the use of tools with a graphical interface to support interaction with the expert. Finally, we emphasize that the content of the primary studies did not allow us to clearly identify whether the researchers implemented interactions with experts in a synchronous or asynchronous way.

- RQ-4** How has expert supervision been evaluated considering the clustering outcomes? And what effects does this expert supervision have on the clustering process and in the clustering outcomes? To analyze the context of this research question, we proposed two categorizations to organize the information extracted from the primary studies. Considering the first categorization, which refers to evaluation strategies, we found that most studies have been dedicated to evaluating the quality of clustering, based mainly on well-established measures in the area of data analysis, although some evaluations also occur through visual inspection of results. In addition, we found that few studies have been devoted to assessing the adherence of clustering to the expectations of the specialist and the effect of supervision on the clustering process. The second categorization, which refers to measures of evaluation of clustering results, allowed us to observe a low interest of interactive clustering researchers in using internal clustering validation indexes. Regarding the gains that expert supervision can bring to clustering results, we found that less than half of the studies compared the results produced by interactive clustering with results produced by classical clustering (i.e., without supervision). Although no study has mentioned failures with the application of interactive clustering, the lack of evidence on differences between results obtained with and without expert supervision does not contribute to building a motivating scenario for interactive clustering. Thus, we highlight three issues in interactive clustering that should be further explored in order to discuss whether expert supervision should in fact be included in the clustering process: (a) evaluating compliance with the expert's expectations and the consequences, possibly harmful, of imposing the expert's knowledge in the clustering process; (b) exploring the use of internal clustering validation indexes in interactive clustering, with the potential development of new indexes or adapted indexes; and (c) comparing results obtained with and without specialist supervision.
- RQ-5** How has the expert's effort been measured? The measurement of the expert's effort is virtually nonexistent in interactive clustering research. Of the few studies that use some strategy to measure the expert's effort, only four of them involve human experts in the experiments conducted. The other studies present some evaluation

strategies but applied only on simulator systems. This is a major research gap in interactive clustering since it is not possible to make a robust analysis of the cost-benefit ratio of inserting humans in the clustering loop without this type of measurement.

7 Conclusion

This scoping review presents an analysis of 50 primary studies, published in the area of interactive clustering, between 2007 and 2018, in the light of five research questions. The analysis of the studies allowed to:

- Draw an overview about interactive clustering, in which we highlight the observation of a trend of growth and maturation of the publications in the area, the attention paid to the development of strategies for visualizing clustering results and the propensity for the application of interactive clustering in the text-mining and image analysis domains.
- Develop a quality analysis of the papers and the research reported therein, which identified that, although the research developed is showing promising results, motivating the continuation of investigations in the area, the area is still immature regarding verifying the validity of results using statistical tests and identifying the actual cost-benefit ratio that the insertion of experts brings to the clustering process.
- Answer the research question RQ-1 in all its aspects; answer the research questions RQ-2, RQ-3 and RQ-4 partially, because of insufficient information in the primary studies analyzed to support assertive responses in some aspects we hoped to clarify; and find out that the target of the research question RQ-5 is not adequately addressed in the primary studies and hence cannot be adequately answered.
- Systematize the content of primary studies in a set of tables that relate the characteristics of research in interactive clustering with each primary study. These tables represent a valuable source of knowledge for researchers who are starting their studies in the area of interactive clustering, who want to position their research in relation to what has already been developed in the area or who are looking for research gaps in which they are able to contribute.

This scoping review is not free from threats to its validity. The subjectivity inherent in the analysis we carried out can affect the results obtained in the selection of primary studies, in the extraction of data from such studies and in the creation and filling in of the categories used to organize the information. This subjectivity can generate different results in attempts to reproduce this work. Part of the harmful effect of these differences can be minimized by consulting the details provided by the tables presented in "Appendix 1".

Acknowledgements This study was financed by: the Coordination of Superior Level Staff Improvement, CAPES-Brazil (CAPES)—Finance Code 001; the São Paulo Research Foundation—FAPESP-Brazil—Process Numbers 2017/26487-4 and 2017/26491-1.

Appendix 1: Tabular information on the primary studies

See Tables 9, 10, 11, 12, 13, 14, 15, 16, 17 and 18.

Table 9 Primary studies—published in *journals*

Author (Year)	Title
Arin et al. (2018)	I-TWEC: interactive clustering tool for Twitter
Cavallo and Demiralp (2018)	Clustrophile 2: guided visual clustering analysis
Nourashrafeddin et al. (2018)	A visual approach for interactive keyterm-based clustering
Sacha et al. (2018)	SOMFlow: guided exploratory cluster analysis with self-organizing maps and analytic provenance
Awasthi et al. (2017)	Local algorithms for interactive clustering
Lei et al. (2017)	Interactive k-means clustering method based on user behavior for different analysis target in medicine
Boudjeloud-Assala et al. (2016)	Interactive and iterative visual clustering
Correa et al. (2015)	Interactive textual feature selection for consensus clustering
Hu et al. (2014)	Interactive document clustering with feature supervision through reweighting
Lai et al. (2014)	A new interactive semi-supervised clustering model for large image database indexing
Xiong et al. (2014)	Active learning of constraints for semi-supervised clustering
Zhang et al. (2014)	Context-assisted face clustering framework with human-in-the-loop
Mukhopadhyay et al. (2013)	An interactive approach to multiobjective clustering of gene expression patterns
Lee et al. (2012)	iVisClustering: an interactive visual document clustering via topic modeling
Wang et al. (2012)	Intelligent photo clustering with user interaction and distance metric learning
Cao et al. (2011)	DICON: Interactive visual analysis of multidimensional clusters
Okabe and Yamada (2011)	An interactive tool for human active learning in constrained clustering
Dasgupta and Ng (2010)	Which clustering do you want? Inducing your ideal clustering with minimal feedback
Zhu et al. (2008)	caBIG TM VISDA: modeling, visualization, and discovery for cluster analysis of genomic data

Table 10 Primary studies—published in *conferences*

Author (Year)	Title
Mai et al. (2018)	Scalable active constrained clustering for temporal data
Sherkat et al. (2018)	Interactive document clustering revisited: a visual analytics approach
Coden et al. (2017)	A method to accelerate human in the loop clustering
Emamjomeh-Zadeh and Kempe (2017)	A general framework for robust interactive learning
Ferrero et al. (2017)	InToEventS: an interactive toolkit for discovering and building event schemas
Chang et al. (2016)	AppGrouper: knowledge-based interactive clustering tool for app search results
Mauder et al. (2016)	GMMbuilder-user-driven discovery of clustering structure for bioarchaeology
Mukhopadhyay (2016)	Interactive approach to multiobjective genetic fuzzy clustering for satellite image segmentation
Vikram and Dasgupta (2016)	Interactive bayesian hierarchical clustering
Vu et al. (2016)	Towards an approach using metric learning for interactive semi-supervised clustering of images
Xu et al. (2016)	Interactive visual co-cluster analysis of bipartite graphs
Gieseke et al. (2015)	Batch steepest-descent-mildest-ascent for interactive maximum margin clustering
Khodabandeh et al. (2015)	Discovering human interactions in videos with limited data labeling
Lelkes and Reyzin (2015)	Interactive clustering of linear classes and cryptographic lower bounds
Senderovich and Maysuradze (2015)	Interactive coding of responses to open-ended questions in Russian
Geerts and Ndindi (2014)	Interactive correlation clustering
Bruneau and Otjacques (2013)	An interactive, example-based, visual clustering system
Marcacini et al. (2013)	Improving consensus clustering of texts using interactive feature selection
Marcacini et al. (2012)	An active learning approach to frequent itemset-based text clustering
Fredj et al. (2011)	An exploration framework for segmentation parameter spaces
Hu et al. (2011)	Interactive feature selection for document clustering
Dubey et al. (2010)	A cluster-level semi-supervision model for interactive clustering
Guo et al. (2010)	Interactive local clustering operations for high dimensional data in parallel coordinates
Ji et al. (2010)	Semi-automatic photo clustering with distance metric learning
Okabe and Yamada (2010a)	Constrained clustering with interactive similarity learning
Okabe and Yamada (2010b)	An interactive tool for constrained clustering with human sampling
Momma et al. (2009)	Promoting total efficiency in text clustering via iterative and interactive metric learning
Schreck et al. (2009)	Visual cluster analysis of trajectory data with interactive kohonen maps
Balcan and Blum (2008)	Clustering with interactive feedback
desJardins et al. (2007)	Interactive visual clustering
Iorio et al. (2007)	An interactive tool for data visualization and clustering

Table 11 Data types and dependence on data type and context

Primary study	Data types					Dependence		
	Text	Image/video	Signals/Time series	Trust graph	Structured	Nonexistent	Data type	
							Data type	Context
Arin et al. (2018)	x						x	x
Cavallo and Demiralp (2018)					x			
Mai et al. (2018)		x	x					
Nourashrafeddin et al. (2018)	x						x	
Sacha et al. (2018)			x					
Sherkat et al. (2018)	x						x	
Awasthi et al. (2017)	x							
Coden et al. (2017)	x							
Emanjomeh-Zadeh and Kempe (2017)						x		
Ferrero et al. (2017)	x						x	x
Lei et al. (2017)					x			
Boudjeloud-Assala et al. (2016)		x	x					
Chang et al. (2016)	x						x	x
Mauder et al. (2016)					x		x	
Mukhopadhyay (2016)		x						
Vikram and Dasgupta (2016)	x	x			x			
Vu et al. (2016)		x						
Xu et al. (2016)	x							
Correa et al. (2015)	x						x	x
Gieseke et al. (2015)		x					x	
Khodabandeh et al. (2015)		x						
Lelkes and Reyzin (2015)						x		
Senderovich and Maysuradze (2015)	x							
Geerts and Ndindi (2014)				x			x	
Hu et al. (2014)	x							

Table 11 (continued)

Primary study	Data types					Dependence		
	Text	Image/video	Signals/Time series	Trust graph	Structured	Nonexistent	Data type	Context
Lai et al. (2014)		x						
Xiong et al. (2014)		x			x			
Zhang et al. (2014)		x					x	x
Bruneau and Oijacques (2013)			x		x			
Marcacini et al. (2013)	x						x	
Mukhopadhyay et al. (2013)					x			
Lee et al. (2012)	x						x	
Marcacini et al. (2012)	x						x	
Wang et al. (2012)		x					x	
Cao et al. (2011)					x			
Fredj et al. (2011)		x					x	
Hu et al. (2011)	x							
Okabe and Yamada (2011)					x			
Dasgupta and Ng (2010)	x							
Dubey et al. (2010)	x							
Guo et al. (2010)					x			
Ji et al. (2010)		x					x	
Okabe and Yamada (2010a)					x			
Okabe and Yamada (2010b)					x			
Momma et al. (2009)	x							
Schreck et al. (2009)			x					
Balkan and Blum (2008)						x	x	x
Zhu et al. (2008)					x		x	x

Table 11 (continued)

Primary study	Data types					Dependence	
	Text	Image/video	Signals/Time series	Trust graph	Structured	Nonexistent	Context
desJardins et al. (2007)					x		
Iorio et al. (2007)					x		

Table 12 Class of the clustering algorithm and algorithm independence

Primary study	Clustering algorithm classes					Algorithm dependent
	Partitioning	Hierarchical	Model-based	Density-based	Other	
Arrn et al. (2018)	x					x
Cavallo and Demiralp (2018)	x	x		x	x	
Mai et al. (2018)	x					
Nourashrafeddin et al. (2018)	x					
Sacha et al. (2018)			x			x
Sherkat et al. (2018)	x					
Awasthi et al. (2017)		x				x
Coden et al. (2017)	x					
Emanjombeh-Zadeh and Kempe (2017)					x	
Ferrero et al. (2017)		x				x
Lei et al. (2017)	x					x
Boudjeloud-Assala et al. (2016)	x					
Chang et al. (2016)		x				x
Mauder et al. (2016)			x			x
Mukhopadhyay (2016)					x	
Vikram and Dasgupta (2016)		x				x
Vu et al. (2016)	x					x
Xu et al. (2016)	x					
Correa et al. (2015)	x					
Gieseke et al. (2015)					x	x
Khodabandeh et al. (2015)					x	x
Lelkes and Reyzin (2015)						
Senderovich and Maysuradze (2015)	x				x	x
Geerts and Ndingi (2014)					x	
Hu et al. (2014)	x					x

Table 12 (continued)

Primary study	Clustering algorithm classes					Algorithm dependent
	Partitioning	Hierarchical	Model-based	Density-based	Other	
Lai et al. (2014)		x				
Xiong et al. (2014)	x					
Zhang et al. (2014)					x	
Bruneau and Oti Jacques (2013)	x					
Marcacini et al. (2013)	x					
Mukhopadhyay et al. (2013)					x	x
Lee et al. (2012)		x				
Marcacini et al. (2012)	x					
Wang et al. (2012)	x					
Cao et al. (2011)	x					
Fredj et al. (2011)	x					
Hu et al. (2011)	x					
Okabe and Yamada (2011)	x					
Dasgupta and Ng (2010)	x					x
Dubey et al. (2010)	x					
Guo et al. (2010)		x				
Ji et al. (2010)	x					
Okabe and Yamada (2010a)	x					
Okabe and Yamada (2010b)	x					
Momma et al. (2009)	x					
Schreck et al. (2009)			x			x
Balkan and Blum (2008)					x	
Zhu et al. (2008)		x				
desJardins et al. (2007)	x					x

Table 12 (continued)

Primary study	Clustering algorithm classes				Algorithm dependent
	Partitioning	Hierarchical	Model-based	Density-based	
Iorio et al. (2007)	x		x		Other x

Table 13 Strategies for acquiring expert's knowledge

Primary study	Data reassignment	Split/merge requests	Pairwise constraints	Feature selection	Feature ranking	Parameter value setting	Data point selection	Seed selection	Clustering evaluation (ranking)	Cluster refine/lock	Other
Ann et al. (2018)		x									
Cavallo and Demiralp (2018)						x	x				x
Mai et al. (2018)			x								
Nourshafeddin et al. (2018)											x
Sacha et al. (2018)		x									
Sherkat et al. (2018)				x							
Awasthi et al. (2017)		x									
Coden et al. (2017)										x	
Emanjomeh-Zadeh and Kempe (2017)		x									
Ferrero et al. (2017)		x				x	x				
Lei et al. (2017)		x			x						
Boudjeloud-Assala et al. (2016)						x		x			
Chang et al. (2016)		x									
Mauder et al. (2016)								x			x
Mukhopadhyay (2016)									x		
Vikram and Dasgupta (2016)			x								
Vu et al. (2016)		x									
Xu et al. (2016)		x	x								
Correa et al. (2015)					x						
Gieseke et al. (2015)						x					
Khodabandeh et al. (2015)			x								
Lelkes and Reyzin (2015)		x									
Senderovich and Maysuradze (2015)		x					x				x
Geerts and Ndindi (2014)			x			x					

Table 13 (continued)

Primary study	Data reassignment	Split/merge requests	Pairwise constraints	Feature selection	Feature ranking	Parameter value setting	Data point selection	Seed selection	Clustering evaluation (ranking)	Cluster refine/lock	Other
Hu et al. (2014)				x						x	
Lai et al. (2014)	x										
Xiong et al. (2014)			x								
Zhang et al. (2014)		x									
Bruneau and Oijacques (2013)			x								
Marcacini et al. (2013)					x						
Mukhopadhyay et al. (2013)									x		
Lee et al. (2012)	x	x					x				
Marcacini et al. (2012)					x						
Wang et al. (2012)	x	x									
Cao et al. (2011)		x		x							
Fredj et al. (2011)											
Hu et al. (2011)				x			x				
Okabe and Yamada (2011)			x								
Dasgupta and Ng (2010)				x							
Dubey et al. (2010)	x				x						
Guo et al. (2010)			x								
Ji et al. (2010)	x										
Okabe and Yamada (2010a)			x								
Okabe and Yamada (2010b)			x								
Momma et al. (2009)		x			x						
Schreck et al. (2009)								x			
Balcan and Blum (2008)		x									
Zhu et al. (2008)								x			

Table 13 (continued)

Primary study	Data reassignment	Split/merge requests	Pairwise constraints	Feature selection	Feature ranking	Parameter value setting	Data point selection	Seed selection	Clustering evaluation (ranking)	Cluster refine/lock	Other
desJardins et al. (2007)	x										
Iorio et al. (2007)	x										

Table 14 Strategies for inserting the expert's knowledge

Primary study	Similarity metric teaming	Split/merge clustering	Pairwise constrained clustering	Seed selection	Feature selection	Parameter value setting	Data point selection	Data reassignment	Feature enrichment	Feature reweighting	Other
Ann et al. (2018)		x									
Cavallo and Demiralp (2018)						x	x				x
Mai et al. (2018)			x								
Nourashrafeddin et al. (2018)				x							
Sacha et al. (2018)		x					x				
Sherkat et al. (2018)				x							
Awasthi et al. (2017)		x									
Coden et al. (2017)	x										
Emanjomeh-Zadeh and Kempe (2017)		x									
Ferrero et al. (2017)		x			x	x	x				
Lei et al. (2017)										x	
Boudjeloud-Assala et al. (2016)				x		x					
Chang et al. (2016)								x			
Mauder et al. (2016)				x							
Mukhopadhyay (2016)	x										
Vikram and Dasgupta (2016)	x										
Vu et al. (2016)	x										
Xu et al. (2016)	x										
Correa et al. (2015)									x		
Gieseke et al. (2015)						x					
Khodabandeh et al. (2015)			x								
Lelkes and Reyzin (2015)		x					x				
Senderovich and Maysuradze (2015)								x			
Geerts and Ndindi (2014)			x			x					

Table 14 (continued)

Primary study	Similarity metric team- ing	Split/ merge clustering	Pairwise constrained clustering	Seed selection	Feature selec- tion	Parameter value set- ting	Data point selection	Data reassign- ment	Feature enrich- ment	Feature reweight- ing	Other
Hu et al. (2014)					x					x	
Lai et al. (2014)	x										
Xiong et al. (2014)	x										
Zhang et al. (2014)		x									
Bruneau and Ojiaques (2013)	x										
Marcacini et al. (2013)									x		
Mukhopadhyay et al. (2013)	x										
Lee et al. (2012)		x					x				
Marcacini et al. (2012)											
Wang et al. (2012)	x								x		
Cao et al. (2011)		x			x						
Fredj et al. (2011)							x				
Hu et al. (2011)					x						
Okabe and Yamada (2011)	x										
Dasgupta and Ng (2010)											
Dubey et al. (2010)					x						
Guo et al. (2010)			x								
Ji et al. (2010)	x		x								
Okabe and Yamada (2010a)	x										
Okabe and Yamada (2010b)	x										
Momma et al. (2009)	x										
Schreck et al. (2009)				x							
Balkan and Blum (2008)		x									
Zhu et al. (2008)				x							

Table 14 (continued)

Primary study	Similarity metric learn- ing	Split/ merge clustering	Pairwise constrained clustering	Seed selection	Feature selec- tion	Parameter value set- ting	Data point selection	Data reassign- ment	Feature enrich- ment	Feature reweight- ing	Other
desJardins et al. (2007)			x					x			
Iorio et al. (2007)											

Table 15 Strategies for presenting clustering information

Primary study	Data points	Features	Similarity	All	Selected	Cluster characterization	Not specified	Abstractions or projections	Graphical interface tool	Complexity analysis
Ann et al. (2018)	x				x				x	
Cavallo and Demiralp (2018)	x	x		x		x		x	x	
Mai et al. (2018)	x				x					x
Nourshafeddin et al. (2018)	x	x		x	x	x		x	x	
Sacha et al. (2018)			x	x				x	x	
Sherkat et al. (2018)	x	x		x				x	x	
Awasthi et al. (2017)							x			
Coden et al. (2017)							x			
Emanjomeh-Zadeh and Kempe (2017)							x			x
Ferrero et al. (2017)							x		x	
Lei et al. (2017)			x	x			x			
Boudjeloud-Assala et al. (2016)	x			x				x	x	
Chang et al. (2016)		x			x				x	
Mauder et al. (2016)	x			x				x	x	
Mukhopadhyay (2016)	x			x					x	
Vikram and Dasgupta (2016)	x				x					
Vu et al. (2016)	x			x						
Xu et al. (2016)			x	x				x	x	
Correa et al. (2015)		x			x				x	
Gieseke et al. (2015)	x			x					x	x
Khodabandeh et al. (2015)	x			x						
Lelkes and Reyzin (2015)							x			x
Senderovich and Maysuradze (2015)	x			x						
Geerts and Ndindi (2014)							x			
Hu et al. (2014)		x			x					x

Table 15 (continued)

Primary study	Data points	Features	Similarity	All	Selected	Cluster char- acterization	Not specified	Abstractions or projections	Graphical interface tool	Com- plexity analysis
Lai et al. (2014)	x			x				x	x	
Xiong et al. (2014)							x			
Zhang et al. (2014)	x			x						
Bruneau and Oijacques (2013)	x			x				x		x
Marcacini et al. (2013)		x			x				x	
Mukhopadhyay et al. (2013)	x			x					x	
Lee et al. (2012)	x	x		x		x			x	
Marcacini et al. (2012)		x			x				x	
Wang et al. (2012)	x			x						x
Cao et al. (2011)	x			x				x	x	
Fredj et al. (2011)	x			x						
Hu et al. (2011)		x			x					
Okabe and Yamada (2011)	x			x				x	x	
Dasgupta and Ng (2010)		x			x					
Dubey et al. (2010)							x			
Guo et al. (2010)		x		x					x	
Ji et al. (2010)	x			x						
Okabe and Yamada (2010a)	x			x				x	x	
Okabe and Yamada (2010b)	x			x				x	x	
Momma et al. (2009)		x				x			x	
Schreck et al. (2009)	x			x					x	
Balkan and Blum (2008)							x			x
Zhu et al. (2008)	x			x				x	x	
desJardins et al. (2007)	x			x				x		

Table 15 (continued)

Primary study	Data points	Features	Similarity	All	Selected	Cluster char-acterization	Not specified	Abstractions or projections	Graphical interface tool	Com-plexity analysis
Iorio et al. (2007)	x			x				x	x	

Table 16 Evaluation strategies

Primary study	Clustering quality measures	Visual analysis	Survey	Process running cost	Algorithm's running cost	Expert effort	Possible experts' choice	None	Comparison
Ann et al. (2018)	x				x				x
Cavallo and Demiralp (2018)									x
Mai et al. (2018)	x		x		x		x		x
Nourashrafeddin et al. (2018)	x								x
Sacha et al. (2018)		x							
Sherkat et al. (2018)	x								x
Awasthi et al. (2017)	x								
Coden et al. (2017)	x								
Emanjomeh-Zadeh and Kempe (2017)								x	
Ferrero et al. (2017)		x							
Lei et al. (2017)	x								x
Boudjeloud-Assala et al. (2016)	x								x
Chang et al. (2016)	x		x						x
Mauder et al. (2016)								x	
Mukhopadhyay (2016)	x								
Vikram and Dasgupta (2016)	x								
Vu et al. (2016)	x								
Xu et al. (2016)			x						x
Correa et al. (2015)	x								x
Gieseke et al. (2015)		x							x
Khodabandeh et al. (2015)	x								x
Lelkes and Reyzin (2015)									
Senderovich and Maysuradze (2015)	x		x					x	
Geerts and Ndindi (2014)				x					
Hu et al. (2014)	x					x			

Table 16 (continued)

Primary study	Clustering quality measures	Visual analysis	Survey	Process running cost	Algorithm's running cost	Expert effort	Possible experts' choice	None	Comparison
Lai et al. (2014)	x				x				x
Xiong et al. (2014)	x			x					
Zhang et al. (2014)	x			x					x
Bruneau and Oijacques (2013)	x								
Marcacini et al. (2013)	x								x
Mukhopadhyay et al. (2013)	x								x
Lee et al. (2012)		x							x
Marcacini et al. (2012)	x								x
Wang et al. (2012)	x								x
Cao et al. (2011)			x						
Fredj et al. (2011)									
Hu et al. (2011)	x	x				x			x
Okabe and Yamada (2011)	x	x							
Dasgupta and Ng (2010)	x								
Dubey et al. (2010)	x								x
Guo et al. (2010)	x	x							
Ji et al. (2010)	x								x
Okabe and Yamada (2010a)	x								
Okabe and Yamada (2010b)	x	x							
Momma et al. (2009)	x								
Schreck et al. (2009)								x	
Balkan and Blum (2008)	x								x
Zhu et al. (2008)	x								
desJardins et al. (2007)	x								

Table 16 (continued)

Primary study	Clustering quality measures	Visual analysis	Survey	Process running cost	Algorithm's running cost	Expert effort	Possible experts' choice	None	Comparison
Iorio et al. (2007)	x								

Table 17 Clustering quality metrics

Primary study	Mutual Informa- tion	F-score	Rand Index	Accu- racy	Jac- card	Recall	Preci- sion	Purity	Sil- hou- ette	V-meas- ure	L-index	Confu- sion matrix	BHI	Ran- dom score	Homo- geneity	Rog- ers- Tani- moto	Person- alized	Oth- ers	None
Arm et al. (2018)		x				x	x										x		
Cavallo and Demiralp (2018)									x									x	
Mai et al. (2018)	x																		
Nourashrafed- din et al. (2018)	x	x																	
Sacha et al. (2018)																			x
Sherkat et al. (2018)	x								x				x	x					
Awasthi et al. (2017)	x																		
Coden et al. (2017)			x																
Emanjomeh- Zadeh and Kempe (2017)																			x
Ferrero et al. (2017)																			x
Lei et al. (2017)																			
Boudjeloud- Assala et al. (2016)			x		x												x		

Table 17 (continued)

Primary study	Mutual Information	F-score	Rand Index	Accuracy	Jaccard	Recall	Precision	Purity	Silhouette	V-measure	L-index	Confusion matrix	BHI	Rand score	Homogeneity	Robustness-Tanimoto	Personalized	Others	None
Chang et al. (2016)																		x	
Mauder et al. (2016)																			x
Mukhopadhyay (2016)											x								
Vikram and Dasgupta (2016)																		x	
Vu et al. (2016)	x																		
Xu et al. (2016)					x													x	
Correa et al. (2015)	x	x																	
Gieseke et al. (2015)																			x
Khodabandeh et al. (2015)								x				x							
Lelkes and Reyzin (2015)																			x
Senderovich and May-suradze (2015)		x																	
Geerts and Ndindi (2014)																	x		

Table 17 (continued)

Primary study	Mutual Information	F-score	Rand Index	Accuracy	Jaccard	Recall	Precision	Purity	Silhouette	V-measure	L-index	Confusion matrix	BHI	Rand score	Homogeneity	Robustness-Tanimoto	Personalized	Others	None
Hu et al. (2014)	x			x	x														
Lai et al. (2014)	x																		
Xiong et al. (2014)	x	x																	
Zhang et al. (2014)						x	x												
Bruneau and Ojiaques (2013)								x											
Marcacini et al. (2013)	x	x																	
Mukhopadhyay et al. (2013)													x						
Lee et al. (2012)																		x	
Marcacini et al. (2012)	x	x																	
Wang et al. (2012)		x																	
Cao et al. (2011)																		x	
Fredj et al. (2011)																		x	
Hu et al. (2011)	x		x																

Table 17 (continued)

Primary study	Mutual Informa- tion	F-score	Rand Index	Accu- racy	Jac- card	Recall	Preci- sion	Purity	Sil- hou- ette	V-meas- ure	L-index	Confu- sion matrix	BHI	Ran- dom score	Homo- geneity	Rog- ers- Tani- moto	Person- alized	Oth- ers	None
Okabe and Yamada (2011)	x																		
Dasgupta and Ng (2010)			x	x															
Dubey et al. (2010)	x	x	x																
Guo et al. (2010)				x															
Ji et al. (2010)		x																	
Okabe and Yamada (2010a)	x																		
Okabe and Yamada (2010b)	x																		
Momma et al. (2009)	x																		
Schreck et al. (2009)																		x	
Balkan and Blum (2008)																		x	
Zhu et al. (2008)				x															
desJardins et al. (2007)			x																
Iorio et al. (2007)																	x		

Table 18 Expert's effort discussions

Primary study	Is expert's effort measured?	Is the expert human?
Arin et al. (2018)		
Cavallo and Demiralp (2018)		x
Mai et al. (2018)	x	
Nourashrafeddin et al. (2018)	x	x
Sacha et al. (2018)		x
Sherkat et al. (2018)	x	x
Awasthi et al. (2017)		
Coden et al. (2017)	x	
Emamjomeh-Zadeh and Kempe (2017)		
Ferrero et al. (2017)		
Lei et al. (2017)		x
Boudjeloud-Assala et al. (2016)		x
Chang et al. (2016)		x
Mauder et al. (2016)		
Mukhopadhyay (2016)		
Vikram and Dasgupta (2016)		
Vu et al. (2016)		
Xu et al. (2016)		x
Correa et al. (2015)	x	
Gieseke et al. (2015)		x
Khodabandeh et al. (2015)	x	
Lelkes and Reyzin (2015)		
Senderovich and Maysuradze (2015)	x	x
Geerts and Ndindi (2014)		
Hu et al. (2014)	x	
Lai et al. (2014)	x	
Xiong et al. (2014)	x	
Zhang et al. (2014)		
Bruneau and Otjacques (2013)		
Marcacini et al. (2013)	x	
Mukhopadhyay et al. (2013)		x
Lee et al. (2012)		x
Marcacini et al. (2012)	x	
Wang et al. (2012)		x
Cao et al. (2011)		x
Fredj et al. (2011)		x
Hu et al. (2011)	x	
Okabe and Yamada (2011)	x	
Dasgupta and Ng (2010)		x
Dubey et al. (2010)		x
Guo et al. (2010)		x
Ji et al. (2010)	x	x
Okabe and Yamada (2010a)		
Okabe and Yamada (2010b)		

Table 18 (continued)

Primary study	Is expert's effort measured?	Is the expert human?
Momma et al. (2009)		x
Schreck et al. (2009)		
Balcan and Blum (2008)		
Zhu et al. (2008)		
desJardins et al. (2007)		
Iorio et al. (2007)		

References

- Achlioptas D, McSherry F (2005) On spectral learning of mixtures of distributions. In: Proceedings of the 18th annual conference on learning theory. Springer-Verlag, Berlin, Germany, COLT'05, pp 458–469
- Alagambigai P, Thangavel K, Visalakshi NK (2008) Interactive clustering in distributed environment. In: First international conference on emerging trends in engineering and technology, pp 1200–1205
- Arin İ, Erpam MK, Saygin Y (2018) I-TWEC: interactive clustering tool for twitter. *Expert Syst Appl* 96:1–13
- Awasthi P, Balcan M, Voevodski K (2017) Local algorithms for interactive clustering. *J Mach Learn Res* 18:1–35
- Balcan MF, Blum A (2008) Clustering with interactive feedback. In: Proceedings of the 19th international conference on algorithmic learning theory (ALT), vol 5254, pp 316–328
- Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S (2016) Pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics* 32(18):2817–2823
- Belkin M, Sinha K (2010) Polynomial learning of distribution families. In: Proceedings of the IEEE 51st annual symposium on foundations of computer science, IEEE computer society, Washington, DC, USA, FOCS '10, pp 103–112
- Biolchini J, Mian PG, Natali ACC, Travassos GH (2005) Systematic review in software engineering. Technical Report TR—ES 679 / 05, Systems Engineering and Computer Science Department, Universidade Federal do Rio de Janeiro, Rio de Janeiro
- Borg I, Groenen P (1997) Modern multidimensional scaling. Theory and applications. Springer, Germany
- Boudjeloud-Assala L, Pinheiro P, Blansch   A, Tamisier T, Otjacques B (2016) Interactive and iterative visual clustering. *Inf Vis* 15(3):181–197
- Brubaker SC, Vempala S (2008) Isotropic PCA and affine-invariant clustering. In: Proceedings of the 49th annual IEEE symposium on foundations of computer science, IEEE computer society, Washington, DC, USA, FOCS '08, pp 551–560
- Bruneau P, Otjacques B (2013) An interactive, example-based, visual clustering system. In: Proceedings of the international conference on information visualisation
- Cao N, Gotz D, Sun J, Qu H (2011) Dicon: interactive visual analysis of multidimensional clusters. *IEEE Trans Vis Comput Graph* 17(12):2581–2590
- Cavallo M, Demiralp    (2018) Clustrophile 2: guided visual clustering analysis. *IEEE Trans Vis Comput Graph* 25:267–276
- Chang S, Dai P, Hong L, Sheng C, Zhang T, Chi EH (2016) AppGrouper: knowledge-based interactive clustering tool for app search results. In: Proceedings of the 21st international conference on intelligent user interfaces, ACM, New York, NY, USA, IUI '16, pp 348–358
- Chapelle O, Schkopf B, Zien A (2010) Semi-supervised learning, 1st edn. The MIT Press, Cambridge
- Chaudhuri K, Dasgupta S (2010) Rates of convergence for the cluster tree. In: Lafferty JD, Williams CKI, Shawe-Taylor J, Zemel RS, Culotta A (eds) *Advances in neural information processing systems*, vol 23. Curran Associates Inc., New York, pp 343–351
- Coden A, Danilevsky M, Gruhl D, Kato L, Nagarajan M (2017) A method to accelerate human in the loop clustering. In: Proceedings of the 17th SIAM international conference on data mining, society for industrial and applied mathematics, Philadelphia, Pennsylvania, United States, pp 237–245

- Cohn D (2010) Active learning. In: Sammut C, Webb GI (eds) *Encyclopedia of machine learning*. Springer, Boston, MA, pp 10–14. https://doi.org/10.1007/978-0-387-30164-8_6
- Correa G, Marcacini R, Hruschka E, Rezende S (2015) Interactive textual feature selection for consensus clustering. *Pattern Recognit Lett* 52:25–31
- Dasgupta S, Ng V (2010) Which clustering do you want? Inducing your ideal clustering with minimal feedback. *J Artif Intell Res* 39(1):581–632
- desJardins M, MacGlashan J, Ferraioli J (2007) Interactive visual clustering. In: 12th International conference on intelligent user interfaces (IUI), pp 361–364
- Dubey A, Bhattacharya I, Godbole S (2010) A cluster-level semi-supervision model for interactive clustering. In: *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* 6321 LNAI (PART 1), pp 409–424
- Dudley JJ, Kristensson PO (2018) A review of user interface design for interactive machine learning. *ACM Trans Interact Intell Syst* 8(2):1–37
- Dyba T, Dingsoyr T, Hanssen GK (2007) Applying systematic reviews to diverse study types: an experience report. In: *First international symposium on empirical software engineering and measurement (ESEM 2007)*, pp 225–234
- Emamjomeh-Zadeh E, Kempe D (2017) A general framework for robust interactive learning. In: *Conference on neural information processing systems, NIPS, 2017*
- Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the 2nd international conference on knowledge discovery and data mining*, pp 226–231. AAAI Press
- Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds) (1996) *Advances in knowledge discovery and data mining*. American Association for Artificial Intelligence, Menlo Park
- Ferrero G, Primadhanty A, Quattoni A (2017) Intoevents: an interactive toolkit for discovering and building event schemas. In: *Proceedings of the software demonstrations of the 15th conference of the European chapter of the association for computational linguistics*, pp 104–107. Association for Computational Linguistics
- Fredj SB, Glatard T, Casta C, Clarysse P (2011) An exploration framework for segmentation parameter spaces. *IEEE Int Conf Image Process* 18:2257–2260
- Geerts F, Ndindi R (2014) Interactive correlation clustering. In: *International conference on data science and advanced analytics (DSAA)*, pp 170–176
- Gharehchopogh FS (2010) Approach and review of user oriented interactive data mining. In: *Proceedings of the 4th international conference on application of information and communication technologies*, pp 1–4
- Gieseke F, Pahikkala T, Heskes T (2015) Batch steepest-descent-mildest-ascent for interactive maximum margin clustering. *Lect Notes Comput Sci* 9385:95–107
- Guo P, Xiao H, Wang Z, Yuan X (2010) Interactive local clustering operations for high dimensional data in parallel coordinates. In: *IEEE Pacific visualization symposium (PacificVis)*, pp 97–104
- Halkidi M, Batistakis Y, Vazirgiannis M (2001) On clustering validation techniques. *J Intell Inf Syst* 17(2):107–145
- Han J, Kamber M, Pei J (2011) *Data mining: concepts and techniques*, 3rd edn. Morgan Kaufmann Publishers Inc., San Francisco
- Holzinger A, Dehmer M, Jurisica I (2014) Knowledge discovery and interactive data mining in bioinformatics—state-of-the-art, future challenges and research directions. *BMC Bioinform* 15(S–6):11
- Hu Y, Milios EE, Blustein J (2011) Interactive feature selection for document clustering. In: *Proceedings of the ACM symposium on applied computing*, pp 1143–1150
- Hu Y, Milios EE, Blustein J (2014) Interactive document clustering with feature supervision through reweighting. *Intell Data Anal* 18:561–581
- Iorio F, Miele G, Napolitano F, Raiconi G, Tagliaferri R (2007) An interactive tool for data visualization and clustering. In: *Knowledge-based intelligent information and engineering systems: 11th international conference (KES), XVII Italian workshop on neural networks*, pp 870–877
- Jang S, Elmqvist N, Ramani K (2014) GestureAnalyzer: Visual analytics for pattern analysis of mid-air hand gestures. In: *Proceedings of the 2nd ACM symposium on spatial user interaction*, ACM, New York, NY, USA, pp 30–39
- Ji D, Wang M, Tian Q, X-SHua (2010) Semi-automatic photo clustering with distance metric learning. In: *Proceedings of SPIE—the international society for optical engineering*, vol 7744
- Kalai AT, Moitra A, Valiant G (2010) Efficiently learning mixtures of two gaussians. In: *Proceedings of the 42nd ACM symposium on theory of computing*, ACM, New York, NY, USA, STOC '10, pp 553–562

- Khodabandeh M, Vahdat A, Zhou GT, Hajimirsadeghi H, Roshtkhari M, Mori G, Se S (2015) Discovering human interactions in videos with limited data labeling. In: IEEE computer society conference on computer vision and pattern recognition workshops, IEEE, New Jersey, USA, vol 2015-October, pp 9–18
- Kitchenham B, Charters S (2007) Guidelines for performing systematic literature reviews in software engineering. Technical report, School of Computer Science and Mathematics Keele University, Keele and Department of Computer Science, University of Durham, Durham
- Kohonen T (2000) Self organizing maps, 3rd edn. Springer, Berlin
- Lai HP, Visani M, Boucher A, Ogier J (2014) A new interactive semi-supervised clustering model for large image database indexing. *Pattern Recognit Lett* 37:94–106
- Lee H, Kihm J, Choo J, Stasko J, Park H (2012) iVisClustering: an interactive visual document clustering via topic modeling. *Comput Graph Forum* 31(3pt3):1155–1164
- Lei Y, Yu D, Bin Z, Yang Y (2017) Interactive K-Means clustering method based on user behavior for different analysis target in medicine. *Comput Math Methods Med* 2017:4915828. <https://doi.org/10.1155/2017/4915828>
- Lelkes AD, Reyzin L (2015) Interactive clustering of linear classes and cryptographic lower bounds. *Algorithmic learning theory*. Springer, Cham, pp 165–176
- Mai ST, Amer-Yahia S, Douzal A, Nguyen KT, Nguyen AD (2018) Scalable active constrained clustering for temporal data. In: 23rd International conference on database systems for advanced applications (DASFAA), pp 566–582
- Marcacini RM, Correa GN, Rezende SO (2012) An active learning approach to frequent itemset-based text clustering. In: Proceedings of the international conference on pattern recognition, IEEE, USA, pp 3529–3532
- Marcacini RM, Domingues MA, Rezende SO (2013) Improving consensus clustering of texts using interactive feature selection. In: Proceedings of the 22nd international conference on world wide web, ACM, USA, WWW'13 Companion, pp 237–238
- Marrero M, Urbano J (2018) A semi-automatic and low-cost method to learn patterns for named entity recognition. *Nat Lang Eng* 24(1):39–75
- Mauder M, Bobkova Y, Ntoutsis E (2016) GMMbuilder-user-driven discovery of clustering structure for bio-archaeology. In: Lecture notes in computer science 9853 LNCS, pp 8–11
- Mitchell TM (1997) Machine learning. McGraw-Hill, New York
- Moitra A, Valiant G (2010) Settling the polynomial learnability of mixtures of gaussians. In: IEEE 51st annual symposium on foundations of computer science. IEEE, Washington, DC, USA, pp 93–102
- Momma M, Morinaga S, Komura D (2009) Promoting total efficiency in text clustering via iterative and interactive metric learning. In: 9th IEEE international conference on data mining, pp 878–883
- Mukhopadhyay A (2016) Interactive approach to multiobjective genetic fuzzy clustering for satellite image segmentation. In: IEEE Uttar Pradesh section international conference on electrical, computer and electronics engineering (UPCON), pp 630–634
- Mukhopadhyay A, Maulik U, Bandyopadhyay S (2013) An interactive approach to multiobjective clustering of gene expression patterns. *IEEE Trans Biomed Eng* 60(1):35–41
- Nourashrafeddin S, Sherkat E, Minghim R, Milios EE (2018) A visual approach for interactive Keyterm-based clustering. *ACM Trans Interact Intell Syst* 8(1):1–35
- Okabe M, Yamada S (2010a) Constrained clustering with interactive similarity learning. In: Joint 5th international conference on soft computing and intelligent systems and 11th international symposium on advanced intelligent systems. Japan society for fuzzy theory and intelligent informatics, Japan, pp 1295–1300
- Okabe M, Yamada S (2010b) An interactive tool for constrained clustering with human sampling. In: Proceedings of the IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology, IEEE computer society, Washington, DC, USA, WI-IAT '10, vol 3, pp 108–111
- Okabe M, Yamada S (2011) An interactive tool for human active learning in constrained clustering. *J Emerg Technol Web Intell* 3:20–27
- Olsson F (2009) A literature survey of active machine learning in the context of natural language processing. Technical Report T2009:06, Swedish Institute of Computer Science, Kista, Sweden
- Paré G, Trudel MC, Jaana M, Kitsiou S (2015) Synthesizing information systems knowledge: a typology of literature reviews. *Inf Manag* 52:183–199
- Patrick EA, Shen LYL (1971) Interactive use of problem knowledge for clustering and decision making. *IEEE Trans Comput C-20*(2):216–222
- Rinaldo A (2010) Generalized density clustering. *Ann Stat* 38:2678–2722

- Sacha D, Kraus M, Bernard J, Behrisch M, Schreck T, Asano Y, Keim DA (2018) SOMFlow: guided exploratory cluster analysis with self-organizing maps and analytic provenance. *IEEE Trans Vis Comput Graph* 24:120–130
- Schreck T, Bernard J, Landesberger T, Kohlhammer J (2009) Visual cluster analysis of trajectory data with interactive kohonen maps. *Inf Vis* 8:14–29
- Schwenker F, Trentin E (2014) Pattern classification and clustering: a review of partially supervised learning approaches. *Pattern Recognit Lett* 37(1):4–14
- Senderovich N, Maysuradze A (2015) Interactive coding of responses to open-ended questions in Russian Knowledge Engineering and Semantic Web. Springer, Cham, pp 195–209
- Settles B (2009) Active learning literature survey. Technical Report CSTR 1648, University of Wisconsin–Madison, Madison, Wisconsin, USA
- Sherkat E, Nourashrafeddin S, Milios EE, Minghim R (2018) Interactive document clustering revisited: a visual analytics approach. In: 23rd International conference on intelligent user interfaces, ACM, New York, NY, USA, IUI'18, pp 281–292
- Sun L, Wang X (2010) A survey on active learning strategy. *Proc Int Conf Mach Learn Cybern* 1:161–166
- Tubío ÓD, Grimán A, Juzgado NJ (2009) Developing search strategies for detecting relevant experiments. *Empir Softw Eng* 14:513–539
- Ullsch A, Siemon HP (1990) Kohonen's self organizing feature maps for exploratory data analysis. In: Proceedings of the international neural network conference, Dordrecht, Netherlands: Kluwer, Paris, France, pp 305–308
- van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9:2579–2605
- Vikram S, Dasgupta S (2016) Interactive Bayesian hierarchical clustering. In: Proceedings of the 33rd international conference on machine learning, JMLR.org, USA, vol 5, pp 3033–3042
- Vu V, Lai H, Visani M (2016) Towards an approach using metric learning for interactive semi-supervised clustering of images. In: Proceedings of the 8th international conference on knowledge and systems engineering, IEEE, USA, pp 357–362
- Wagstaff KL (2010) Constrained clustering. In: Sammut C, Webb GI (eds) Encyclopedia of machine learning. Springer, Boston, MA, pp 220–221. https://doi.org/10.1007/978-0-387-30164-8_163
- Wang M, Ji D, Tian Q, Hua XS (2012) Intelligent photo clustering with user interaction and distance metric learning. *Pattern Recognit Lett* 33(4):462–470
- Wold S, Esbensen K, Geladi P (1987) Principal component analysis. *Chemom Intell Lab Syst* 2(1):37–52
- Xiong S, Azimi J, Fern XZ (2014) Active learning of constraints for semi-supervised clustering. *IEEE Trans Knowl Data Eng* 26(1):43–54
- Xu P, Cao N, Qu H, Stasko J (2016) Interactive visual co-cluster analysis of bipartite graphs. In: IEEE Pacific visualization symposium (PacificVis), pp 32–39
- Zhang XL, Simpson T, Frecker M, Lesieutre G (2012) Supporting knowledge exploration and discovery in multi-dimensional data with interactive multiscale visualisation. *J Eng Des* 23(1):23–47
- Zhang L, Kalashnikov DV, Mehrotra S (2014) Context-assisted face clustering framework with human-in-the-loop. *Int J Multimed Inf Retr* 3:69–88
- Zhang Y, Luo W, Mack EA, Maciejewski R (2016) Visualizing the impact of geographical variations on multivariate clustering. *Comput Graph Forum* 35:101–110
- Zhu Y, Li H, Miller DJ, Wang Z, Xuan J, Clarke R, Hoffman EP, Wang YJ (2008) caBIG™ VISDA: modeling, visualization, and discovery for cluster analysis of genomic data. *BMC Bioinform* 9:383–383

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.