Int. J. Nonlinear Anal. Appl. 14 (2023) 9, 209–240
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2022.29161.4077

Clustering ensemble selection: A systematic mapping study

Hajar Khalili[a], Mohsen Rabbani[b,*], Ebrahim Akbari[a]

[a]*Department of Computer Engineering, Sari Branch, Islamic Azad University, Sari, Iran*
[b]*Department of Applied Mathematics, Sari Branch, Islamic Azad University, Sari, Iran*

(Communicated by Madjid Eshaghi Gordji)

Abstract

Clustering has emerged as an important tool for data analysis, which can be used to produce high-quality data partitions as well as stronger and more accurate consensus clustering based on basic clustering. Data item labels, which are already known as opposed to classification issues, are unlabeled clusters in unsupervised clustering, which may cause uncertainty in large libraries. Therefore, all clusters produced are not useful for the final clustering solution. To address this challenge, instead of selecting all of them from a subset of variants to combine for the obtainment of the final result, Clustering ensemble selection (CES) was proposed in 2006 by Hadjitodorov. The goal is the selection of a subset of large libraries to produce a smaller cluster offering higher-quality performance. (CES) has been found effective in the improvement of the clustering solutions quality. The current paper conducts a systematic mapping study (SMS) for the analysis and synthetization of the studies formerly conducted on the CES techniques. To this end, 42 prominent publications from the existing literature, published from 2006 to August 2022, were selected to be examined in this article. The analysis results showed that most of the articles have used the NMI measure to evaluate the cluster quality, and the method of valuing the initial parameter has been more commonly used for the generation of diversity. Clustering ensemble selection has not been done on text yet; in addition, the trade-off between diversity and quality (considering both at the same time) can be studied and evaluated in the future.

Keywords: Clustering Ensemble Selection, Diversity, Measure, Consensus Function
2020 MSC: 62H30

1 Introduction

Data analysis is the basis of many computational applications, both in the design phase and as part of their online operations. Depending on the accessability of proper models for the data source,data analysis methods fall into two types, i.e., exploratory and confirmatory. However, a crucial element to form a hypothesis or decision is to group or classify. Measurements are based on being fit with a hypothetical model or natural groupings (clustering) that are revealed through analysis.

Cluster analysis organizes a set of patterns (usually represented as a vector of measurements or a point in multidimensional space) based on similarity to clusters[27]. Cluster analysis has been recognized in the literature as a key approach since it classifies the elements of a dataset regarding their similarity, without the need for any class

*Corresponding author
Email addresses: khalili.hajar1982@yahoo.com (Hajar Khalili), mrabbani@iausari.ac.ir (Mohsen Rabbani), ebrahimakbari30@yahoo.com (Ebrahim Akbari)

Received: October 2022 *Accepted:* December 2022

label information. In addition, clustering techniques are applicable to the analysis of biological data with different characteristics. The challenge of choosing the optimal algorithm and types of clustering methods typically results in conflicting outcomes because of methodological bias and different performance criteria [20, 21].

So far, the most important objective of the groups has been the enhancement of the accuracy and effectiveness of a particular classification or regression. Significant improvements have also been made to a wide range of datasets[44]. Contrary to the classification or regression settings, the literature consists of very few approaches introduced for the combination of multiple clusterings. In the following, the most important exceptions are presented:

- Accurate consensus clustering to design evolutionary trees, leading to solutions with much lower resolution than individual solutions.
- Combining the results of several clusters from a given dataset, in which each solution of the combination is in a common, well-known space, for example, combining multiple sets of cluster centers using k-means. It is obtained with different initial values [11].

The rapid advancement of clustering science and technology has caused clustering to play a key role in different fields, e.g.,image processing, pattern recognition, document clustering, business intelligence, market research, customer recommendations, and data analysis. It is not easy to find a clustering algorithm applicable to all data sets; as a result, the literature is loaded with different clustering algorithms.To solve this problem, the concept of clustering is proposed in 2003 [47].

A consensus of different clustering partitions combines the dataset into a final partition. The result of the clustering set is superior to the single clustering algorithm. The single clustering algorithm, due to its special weakness, leads to an algorithm only for a specific dataset. The clustering consensus combines these clustering algorithms to eliminate the violations of the single clustering algorithm that conforms to more data than clustering and is also noise resistant [49].

The basic algorithm generates consensus members using k-means with different initial values and combines members using cumulative clustering with single, average, complete link. Next, the effect of consensus size on the clustering set is analyzed to find the appropriate consensus size. In addition, the relationship between the diversity and performance of the clustering consensus is examined to guide the selection of consensus members. Finally, the selected clustering set is compared with the traditional clustering consensus based on quality and variety.

The aim of the present systematic mapping study (SMS) is to summarize and integrate the available studies using the following five research questions (RQs):

a. What years have the selected studies been conducted on CES (RQ1)?
b. What is the diversity (RQ2)?
c. How base clusterings are generated in different methods (RQ3)?
d. Which journals have paid more attention to CES (RQ4)?
e. Which measures are worked in CES (RQ5)?

The rest of the paper is structured as follows: Section 2 briefly describes the SMS studies previously conducted on Clustering ensemble selection. Then, Section 3 gives the methodology that describes the methods and materials employed in performing this SMS. Next, Section 4 reports the findings related to each research question. Afterwards, Section 5 discusses the obtained results and presents their implications for the research body. Finally, the last section presents the conclusion and recommends directions for further work in this domain.

2 Related Work

Clustering is a key step to data mining, which seeks to divide data into groups or clusters based on specific similarity criteria. The general purpose of clustering is to place similar data points in a cluster, hence improving the robustness and quality of clustering results. The literature consists of many approaches to solving the set problems [41]. The goal of ensemble clustering is the combination of several clusters for a possibly better and stronger clustering result, which has the advantage of finding bizarre clusters, dealing with noise, and integrating clustering solutions from different sources [53].In general, a clustering set consists of two parts: the first step is to create a diverse set of base clusters; they should be different from each other because the diversity between base clusters helps to improve group performance. The second step is the solution and combination of multiple clusters (e.g., consensus function and the aggregation of multiple clusterings) [34, 1, 22, 24, 23, 6, 2, 58, 42].

Consensus clustering has been reviewed by a number of scholars [16, 51, 9, 54]. Given that members are in unlabeled clustering, not all clustering results can be expected to be useful for the final consensus clustering solution [51, 9].It has recently been shown that better clustering can be achieved by using a subset of clustering members [54]. Recently, it has been proven that a subset of clustering members can be used to achieve better clustering[18]. This approach is termed clustering ensemble selection (CES). The main idea of selecting group clustering to form a cluster group is the selection of a diverse subset of smaller base clusters that perform better than all clustering members[5]. In case of unsupervised clustering, there is not the same external objective function for the measurement of the clustering quality as accuracy.

In the clustering literature, predefined class labels are commonly used as an alternative to the main structure in order to measure the quality of clustering. However, this can not be applied to set selection since supervised information such as class tags cannot be involved in the clustering process[47]. The literature comprises various diversity measures applicable to cluster ensembles[12]. Diversity and quality are considered as two crucial criteria for selecting basic clustering and influencing group performance. Diversity is very important for the success of group clustering because high quality basic clustering affects the performance of the final clustering solution. Variety and quality are shown in CES, which leads to an increase in final results compared to complete sets [14]. The relationship between diversity and quality is unclear. To increase quality, diversity is increased by removing additional base partitions[52]. Figure 1.a shows the clustering according to the input data; Figure 1.b shows the different clusters extracted from the data by a consensus function of clusters of higher quality than figure 1.a; then, in Figure 1.c, higher quality clusterings are produced due to the omission of some clusters.

3 Methodology

The main purpose of an SMS is identifying, counting, and classifying all studies dedicated to an extensive research field. Then, after evaluating and interpreting the findings of the articles, a basic question is answered by combining the obtained results. Survey studies are of great importance because they can give an interesting review to make progress in that area. In addition, SMS can be taken into account as a valuable basis for more accurate systematic review and follow-up. A survey study presents a review of a study area through the identification of the quantity and type of studies that have been published in that field to determine the gaps and research trends, whereas a systematic review employs a more accurate and completely-defined method for the purpose of reviewing the existing literature on a particular topic. In the end, a systematic map widely addresses and analyzes the selected papers and designates the method they use. Figure 2 presents the five significant steps of a systematic survey, which are (1) defining the research questions, (2) searching for pilot studies, (3) screening articles, (4) writing keywords, and (5) extracting data and surveying.

3.1 Research Questions

For the formulation of the research questions in an SMS, a popular approach is the implementation of the PIOC (Population Intervention Outcomes Context) criterion. Research questions prepared using PIOC are structured in four aspects: (a) population; (b) intervention; (c) result; and (d) context. The PIOC characteristics of the research questions are shown in Table 1.

Table 1: Summary of PIOC

Population	Clustering ensemble selection
Intervention	Diversity, Quality
Outcomes	High quality cluster and optimal selected clusterings
Context	The Relationship between quality and diversity

The main purpose of the current SMS study is the identification and evaluation of the articles published between 2006 and August 2022 based on Clustering Ensemble Selection. The five research questions set for this study are given in Table 2, and their motivation and variables were formulated with the aim of achieving a clear attitude toward the subject.

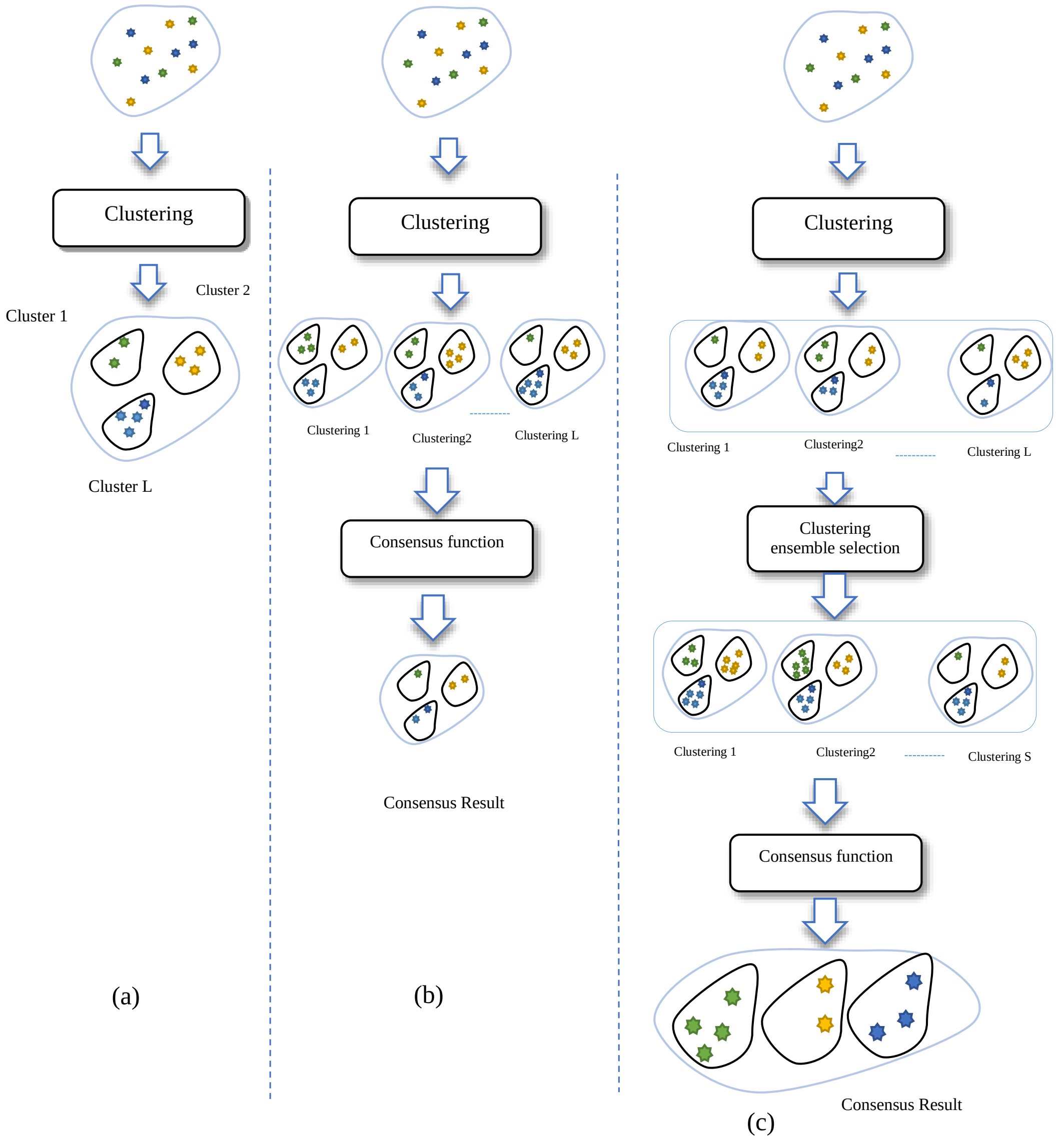

Figure 1: Process of clustering, Clustering Ensemble and Clustering Ensemble Selection Approaches

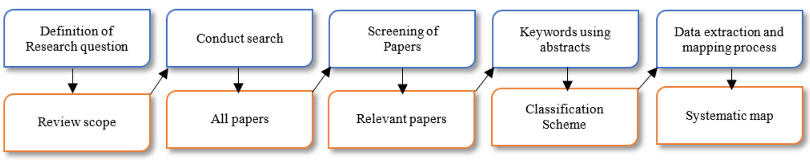

Figure 2: The sms process

Table 2: Research questions

RQ	Research questions	Motivation	Variable
RQ1	What years have selected studies been done on CES?	Specify areas and when efforts have been made in this field.	Research Year
RQ2	What is the diversity?	Because diversity are important in base clustering and consensus result	measures
RQ3	How base clusterings are generated in different methods?	One consensus function on different diversities obtain different consensus function results	Research Methods
RQ4	Which measures are worked in CES?	The effect of measures on quality and diversity	Quality measure and diversity measure
RQ5	Which journal have paid more attention to CES?	Determine which journals are related to the CES	Research Publisher

In general, the aim of an SMS is to conduct pertinent research for the purpose of evaluating the evidence available to deal with RQs. This trend should be strict and impartial and often involves extensive coverage of resources, e.g., online databases and journals. For the minimization of bias and maximization of the number of resources examined, a predefined strategy is needed for the identification of pilot studies, as described in Table 3.

Table 3: Terms obtained from PIOC

Population	Refers to the applied field where we pay attention to CES,
Intervention	Instruments, techniques, methods, and technology to be studied. In this study, we pay attention to Relationship between quality and diversity for CES to improve the quality of clustering .
Outcomes	The results are measurable from studies. In this study, we do not pay attention to the study findings.
Context	It refers to the various strategies that have been used, meaning search terms related to the classification trend.

3.2 Search strategy

Article search is done with two search strategies: manual search and automatic search.

3.2.1 Manual search

In Manual search, articles are extracted from journals and researchers' personal page.

3.2.2 Automatic search

In this article, automatic search was used to extract relevant articles from databases using Start software.The strategy implemented for making the searching terms consists of four steps: 1) the main terms were specified, concerning the research questions (PIOC) (Table 2). 2) The synonym of the words or substitute words for the original terms was identified considering the keywords in the articles related to CES (see Table 3). 3) Boolean OR was used as synonyms of alternative words or abbreviations (see Table 4). 4) Finally, Boolean AND was used with the aim of linking the original terms (see Table 5). To reduce the probability of bias, the search string in this study was performed in all selected databases using a specialized search engine in academic cases, and it was measured to evaluate the completeness of the string as the number of related studies identified. This search string is formed with the help of Boolean logic to ensure the comparison of results between databases. After the experiment, we checked the search string. After defining the search terms, the identification of the related literature began. The current search is done on the basis of four electronic databases: Google Scholar, IEEE, Springer, and Science Direct. These databases were selected considering the prevailing literature on the CES. The details in regard to all pilot studies related to the use of Start software, as the free source bibliography reference administrator, were saved. The "export" feature, which is accessible within many electronic databases, was employed in order to automatically export the details of all pilot studies (e.g., title, author(s), abstract, keywords, publication year, and data source name) to Start.

Table 4: searching for substitute words using BOOLEAN OR.

NO.	Main Subject	Result
1	Clustering Ensemble Selection	(selection clustering ensemble OR clustering ensemble selection OR selective clustering ensemble)
2	Data Mining	(data analysis OR data mining OR information discovery OR knowledge discovery(
3	Diversity	(diversity AND quality)

Table 5: consistency of all possible words using BOOLEAN AND.

Final String
("selection clustering ensemble OR clustering ensemble selection OR selective clustering ensemble ") AND ("data analysis" OR "data mining" OR "information discovery" OR "knowledge discovery") AND ("diversity AND quality ")

After defining the keywords, queries were made. These queries were different for each digital library and had different boundary features depending on the digital library facilities. Digital libraries have specific limitations during searching. For example, some of them are not allowed to use full search strings. Some others should complete these strings with a simple text search. For this reason, separate queries should be made for each library and then the general results of these searches should be obtained based on the proposed main queries. Table 6 shows a set of examples for each digital library.

Table 6: Final String in the Databases

Digital Database	String
Springer	(TITLE-ABS-KEY (" selection clustering ensemble " OR " clustering ensemble selection " OR " selective clustering ensemble ") AND TITLE-ABS-KEY ("data analysis" OR "data mining" OR "information discovery" OR "knowledge discovery") AND TITLE-ABS-KEY ("diversity" OR "quality")) AND (LIMIT-TO (DOCTYPE, "ar")) AND (LIMIT-TO (SUBJAREA, "COMP") OR LIMIT-TO (SUBJAREA, "BIOC") OR LIMIT-TO (SUBJAREA, "ENGI") OR LIMIT-TO (SUBJAREA, "MEDI") OR LIMIT-TO (SUBJAREA, "DECI")) AND (LIMIT-TO (LANGUAGE, "English")) AND (LIMIT-TO (SRCTYPE, "j"))
Science Direct	("selection clustering ensemble " OR " clustering ensemble selection " OR " selective clustering ensemble ") AND ("data analysis" OR "data mining" OR "information discovery" OR "knowledge discovery") AND ("microarray" OR "gene expression") Filters applied: Research articles.
Google scholar, IEEE	((“selection clustering ensemble ”[Title/Abstract] OR ” clustering ensemble selection ”[Title/Abstract] OR ” selective clustering ensemble ”[Title/Abstract]) AND (”data analysis”[Title/Abstract] OR ”data mining”[Title/Abstract] OR ”information discovery”[Title/Abstract] OR ”knowledge discovery”[Title/Abstract])) AND (”diversity”[Title/Abstract] OR ”quality”[Title/Abstract]) Filters applied Journal Article, English, and Humans.

3.3 Study selection

The papers that satisfied at least one of the exclusion criteria (ECs) were left out of this study. On the other hand, those papers that satisfied at least one of the inclusion criteria (ICs) and did not satisfy any ECs were kept. Table 7 describes ICs and ECs applied in this study.

Table 7: Inclusion and Exclusion Criteria for Selecting Articles

IC	Inclusion Criteria (IC)	EC	Exclusion Criteria (EC)
IC1	studies from 2006 to August 2022	EC1	Duplicated studies (only one copy of each study was included)
IC2	studies with CES technique	EC2	studies on supervised or FCM method
IC3	studies in computer science	EC3	Non-English writer papers
IC4	studies published in journal	EC4	short paper (<=5 page)
IC5	primary studies	EC5	secondary studies

The studies were selected in three steps. At step 1 (Planning), Google was used to identify the relevant articles by searching for titles, abstracts, and keywords along with key phrases in various databases for inclusion in the Start software. Then, at step 2 (Selection), the titles, summaries, and keywords were screened for the aim of deciding whether or not to take account of the study. In addition, a review was done on the studies on the basis of the inclusion and exclusion criteria. The texts of these articles were read completely. As a result, at step 3 (Execution), the full text of the pilot studies in the preliminary selection was attained. The full text of each pilot study was read in detail, which is included in the preliminary selection. It was done with the aim of deciding to select or delete that study. The pilot studies included in the final selection are based on the relevant articles that satisfied RQs provided in this SMS. The pilot studies were searched according to the above instructions. First, pilot studies were looked for within the databases. Therefore, a total of 515 studies were obtained from the automatic search. It was done by the Start software in two stages, selection and extraction. The pilot studies were chosen through reading the titles and summaries and using the inclusion and exclusion criteria at the next step. Consequently, 42 studies were chosen for the purpose of this research. Therefore, a total of 42 relevant studies were identified from 4 automatic search sources.In Figure 3 see List of automatic search results in the selected electronic databases, In Figure 4 see the number of automatically selected articles from databases, the purple color is considered as the other resources and Figure 5 shows the process of selecting the articles.

Source	URL	Search Results	papers In the selection stage			papers In the extraction stage		
			Accept	Reject	Duplicate	Accept	Reject	Duplicate
Science Direct	www.sciencedirect.com	32	67	77	371	42	15	10
Springer	https://link.springer.com/	346						
IEEE	https://ieeexplore.ieee.org/	82						
Google Scholar	https://scholar.google.com	52						
Total		515						

Figure 3: List of automatic search results in the selected electronic databases

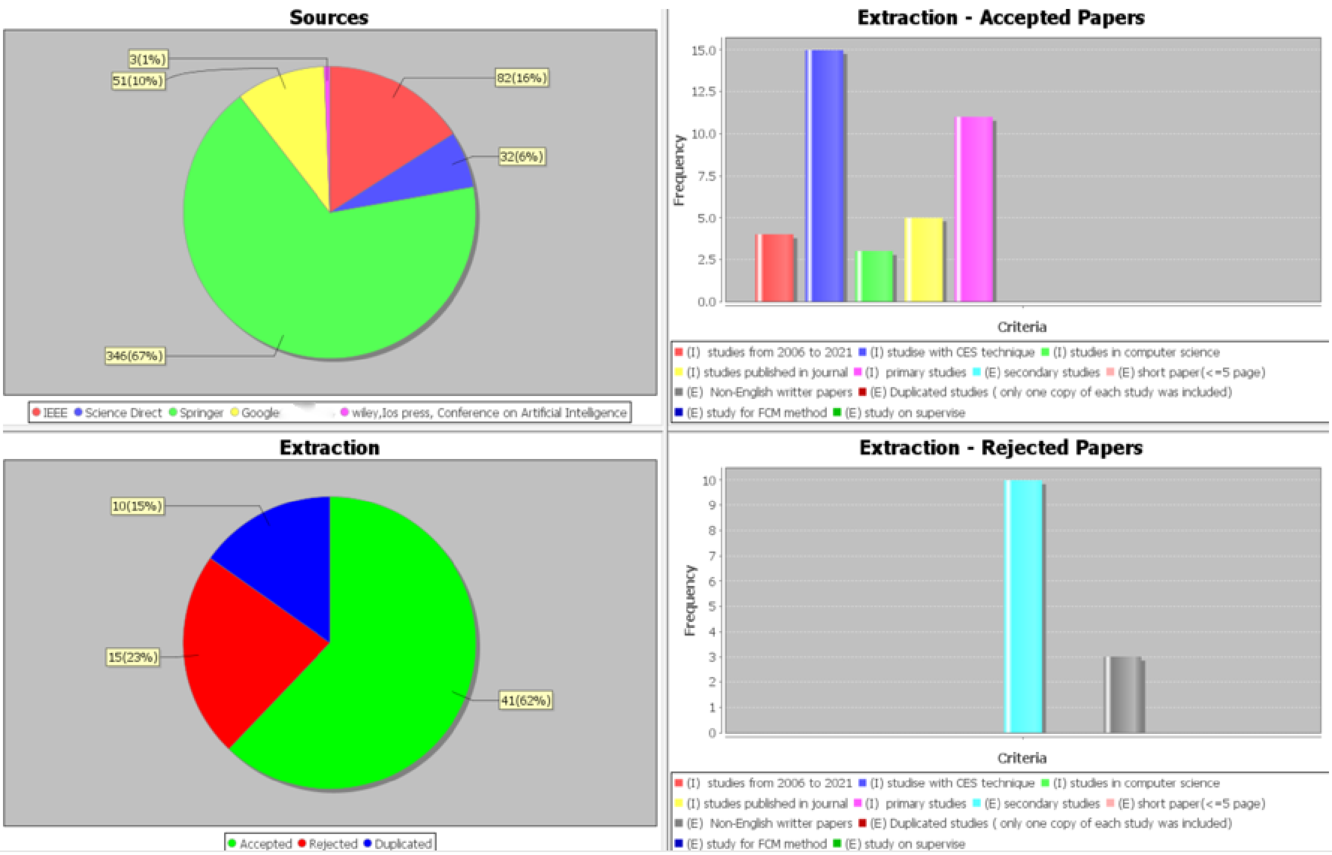

Figure 4: The number of automatically selected articles from databases

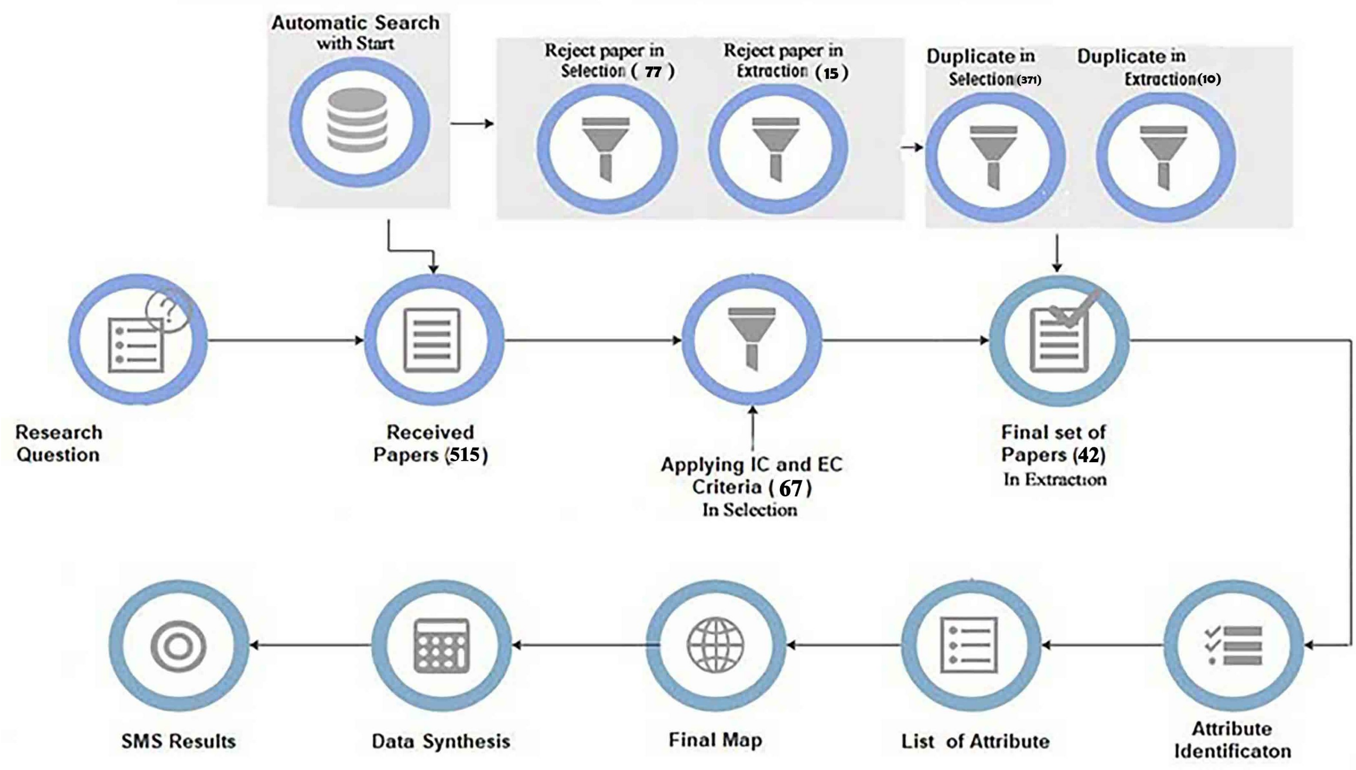

Figure 5: Selected article selection process

3.4 Diversity Generation

Previously-conducted studies have proposed various methods for the creation of diversity or group members, which are listed below. If the clustering quality is improved when using ensemble, they could be of more benefits to users [47]. Stable results of the problem Consensus clustering achieves stable results by calculating the results of basic clustering [17]. The result of clustering composition is better than the basic clustering methods due to its higher strength[47, 8]. Consensus clustering involves the following two methods: (1) diversity, by which multiple clusters are created. Various methods have been proposed to produce diversity, including the following:

a. Valuing the initial parameters: called homogeneous sets, the initial clustering is created by repeatedly performing the clustering algorithm with the k-means technique clustering centers [15].
b. Clustering Algorithms: Using clustering algorithms to generate primary clusters known as heterogeneous sets [48, 7].
c. Different subsets of features: Select features to generate subsets[15, 48, 19].
d. Different subsets of objects: sampling data with or without alternatives[38, 39].
e. Projection to the subspace: Types of one-dimensional and random cuts when throwing objects on the subspaces[48, 7, 19, 38, 39, 12, 55].

And (2) consensus function, in which the multiple clusters produced are merged. Using a number of these approaches, individual clustering diversity is improved[4]. And in the next step, several methods are proposed to combine these multiple clusters[59, 56, 53]. The consensus functions obtained from the composition of the initial clustering are effective in improving the accuracy of the final clustering[45, 13, 43]. The literature includes two criteria of quality and diversity that are applied to group members. The matching index between the two partitions is the basis of this criterion. Normalized reciprocal information (NMI)[47]and adjusted rand index (ARI)[25]are two criteria used by many researchers for diversity and quality assessment between two partitions. For example, Zhong and Gush[60]used NMI to evaluate between clusters, while Kandylas et al.[28]used it in knowledge analysis. In another study, Hadjitodorov et al.[18] used ARI to select each member of the group. Lu et al.[35]proposed a criterion of variety based on covariance. Alizadeh et al.[4]proposed a method in which the selection of clusters was based on diversity and quality.

3.5 Consensus function

The consensus function algorithm combines the members of different groups or clusters in a way to achieve final clusters. that can be divided into voting, paired similarity, feature-based approach and graph-based. The pairwise method creates a correlation matrix in which the similarity between points is the number of times the points are in the same clusters created from the clusters. Hierarchical algorithms such as average-link, single-link, and complete-link are commonly used to combine results using correlation matrices[15]. The voting method is also known as the re-labeling method. Unlike other methods, there is no need to match the labels of the obtained clusters. This method solves the problem of matching between the labels[32].In the feature-based method, the output generated by each clustering algorithm is a classified feature. Clustering algorithms work as new examples on categorized properties.A consensus function is considered as a method that is developed on the basis of the generalized mutual information[50].Formulation of the consensus function is used to solve the problem generated in k-way min-cut hyper graph partitioning [37].On the other hand, the review of the literature shows a challenge in the relationships between diversity and quality and the impact of the two on the group. Strehl and Ghosh [47]proposed three methods of consensus functions: cluster-based similarity algorithm (CSPA), segmentation algorithm (HGPA), and meta-cluster algorithm (MCLA). CSPA creates a pairwise similarity matrix or correlation matrix.

The Hypergraph Segmentation Algorithm (HGPA) function requires different basic clustering. on the other hand, (MCLA) provides more precise solutions to each set.

Table 8 shows the advantages and disadvantages of related clustering ensemble selection. Table 9 compares the CES methods and also shows the different methods used to select clustering sets and different algorithms applied to the generation of basic clustering. in addition, this table compares the articles regarding their use of pairwise, non pairwise, or hybrid approaches based on diversity measurements as well as different consensus functions to generate the final solution.

4 Result

In this section, the results corresponding to the research questions of Table 2 are presented. First, the results of the selection are presented; then, the results of the research questions 1-5.

4.1 RQ1: What years have the selected studies been conducted on CES?

Figure 6 shows the number of the studies selected based on the number/year of studies from 2006 to August 2022. The journal is the source of the 42 selected studies. It is noteworthy that studies on the choice of composite clustering have been started since 2006, and only one study was published in that year by Hedjitodrov, which is considered as the first major work in this field. Additionally, according to Figure 5, in 2015, the most articles (14.2%) were published in the field of composite cluster selection. Then in 2014 and 2018 with 11.9 %, in 2021 with 9.5 %, in 2009, 2019, and 2020 with 7.1 %, in 2011, 2012, 2013, 2016, and 2017 with 4.7% and in 2006, 2008, and 2022 with 2.3 %. The lowest number of surveys was published in 2006 and 2008 with 2.4%

Table 8: Advantages and Disadvantages of related Clustering Ensemble Selection

ID	Journal	Title	Advantages	Disadvantages	Description
S1	Engineering Applications of Artificial Intelligence	Hierarchical cluster ensemble selection	Significant performance improvement compared to ensemble groups	Lack of relationship between diversity and quality in the selection of ensemble members	Using the Hierarchical ensemble Selection method and measuring diversity to examine how diversity and quality affect the final results
S2	Neurocomputing	Cluster ensemble selection with constraints	The AQD2 method has the best performance and has quality and compatibility with diversity.	little research efforts to combine previous background knowledge	Study ensemble clustering and semi-supervised clustering and various techniques for finding high quality solutions
S3	Artificial Intelligence Review	Clustering ensemble selection considering quality and diversity	Using ENMI as the best cluster evaluation and using Average-Linkage algorithm as aggregator along with EEAC and ItoU methods is the best option for consensus function	Consider applying sampling mechanisms and using other rapid metrics to evaluate clusters for the algorithm	assess the association between a cluster and a partition which is called Edited Normalized Mutual Information, ENMI criterion
S4	Data Mining and Knowledge Discovery	Cluster ensemble selection based on relative validity indexes	the impact of the diversity among partitions used for the ensemble	a ground truth (known clustering solution) is not available.	Examining several methods for evaluating and selecting partitions based on relative clustering validity indicators
S5	Pattern Recognition Letters	Bagging-based spectral clustering ensemble selection	Achieve a better clustering solution than traditional clustering methods, especially when the learner is weak.	Expensive and sensitive to scaling parameters and problems of open SC issues and some of its features for individual diversity	Generalization of the selective clustering set algorithm proposed by Azimi and Fern and a new method of selective spectral clustering group (SELSCE)
S6	Soft Computing	Multiple clustering and selecting algorithms with combining strategy for CES	Good performance on most data sets as well as relative to selective clustering algorithms	Study more single CES , research about selection proportion on different data sets, Examine other hybrid strategies	Study the CES problem and propose an MCAS approach considering quality and diversity
S7	Pattern Recognition	Clustering ensemble selection for categorical data based on internal validity indices	Improving the robustness and effectiveness of clustering results by integrating different base clusters based on criteria.	Automatically determines the number of selected base partitions	Selecting a new strategy to improve the performance of set clustering algorithms for classification data namely Sum of Internal Validity Indices with Diversity (SIVID)
S8	International Conference on Neural Information Processing, Springer	Clustering Ensemble Selection with Determinantal Point Processes	Using the DPP method A flexible method for selecting base clusters	Improve the efficiency of DPP clustering sampling	Review of basic clustering selection from a random sampling perspective and propose a clustering selection method with deterministic point processes
S9	ACM Transactions on Knowledge Discovery from Data	Cluster's quality evaluation and selective clustering ensemble	The effect of SME on clustering weighting in a set and DSME in discovering the grouping structure of a dataset	Extend SME to a modified index for chance and size selection	Propose a new criterion for SME and the impact of some SME features on measuring the quality of each cluster in the collection

ID	Journal	Title	Advantages	Disadvantages	Description
S10	International Conference on Fuzzy Systems and Knowledge Discovery, IEEE	Similarity-based spectral clustering ensemble selection	Better clustering performance than traditional methods in the clustering set method when the learner is weak.	Computationally expensive algorithm and sensitivity to scaling parameter during matrix construction	Introducing a new pruning algorithm for unsupervised group learning and a new ensemble method, Selective Spectral Clustering (SELSCE)
S11	Engineering Applications of Artificial Intelligence	A new selection strategy for selective cluster ensemble based on diversity and independency	Improved accuracy of final results compared to other cluster ensemble methods	Use any other metric as weights in WAEC for different clustering solutions	Using an exploratory metric based on code-to-graph conversion in software testing to calculate the independence of the two basic clustering algorithms.
S12	In Proceedings of the 2015 IEEE/ACM International Conference	A multiplex-network based approach for CES	the effectiveness of the proposed CES approach	increasing the use of a set of indicators instead of using a single quality / diversity index.	Introducing a CES approach with the possibility of considering quality and diversity
S13	In 2012 IEEE Ninth International Conference on e-Business Engineering, IEEE	A new selective clustering ensemble algorithm	Significant improvement in clustering performance and algorithm efficiency	Using KMEANS as an alternative to a variety of clustering algorithms in addition to using it as a generation of clustering partitions	Selecting the best reference partition based on the evaluation of clustering validity and presenting a new selection strategy and method of member weight
S14	In Recent Advances of Neural Network Models and Applications, Springer, Cham	A quality-driven ensemble approach to automatic model selection in clustering	more weight to the best-performing (in terms of the selected quality indices) clustering method	Model selection is a major clustering constraint and an inherent problem that cannot be fully answered	the combined use of two different clustering paradigms and their combination by means of an ensemble technique
S15	In 2019 5th International Conference on Big Data and Information Analytics (BigDIA), IEEE.	Selective Ensemble Method Based on Spectral Clustering	Improved spectral clustering performance and results of stable clustering and high clustering accuracy compared to other clustering models	Use any other metric For comparison	Introduction of a set selection method based on spectral clustering
S16	Wuhan University Journal of Natural Sciences	Adaptive spectral clustering ensemble selection via resampling and population-based incremental learning algorithm	Better results compared to traditional clustering methods with the proposed algorithm when the number of component clustering is high	Not all clustering results may be valid, and it is also difficult to access individual clustering diversity, which is a necessity in group learning, if the number of components is large.	Discover a new set method for spectral clustering

ID	Journal	Title	Advantages	Disadvantages	Description
S17	Pattern Recognition	Ensemble Selection with Joint Spectral Clustering and Structural Sparsity	Less sensitive ES-JSS to the type of basic learners, Strong set selection result to test samples using less space	Use stronger self-monitoring learning techniques to select ensemble in unlabeled predictive space, compare performance appraisals	Proposing a new method of static set selection called set selection with common spectral clustering and structural scattering, integration of spectral clustering and structural scattering in a common framework
S18	In International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management (pp. 133-148). Springer, Berlin, Heidelberg	Average cluster consistency for cluster ensemble selection	High quality of the partitions selected by the mentioned measure in comparison with the consensus partitions selected by the other measure.	A method for constructing a cluster and selecting the type of consensus function for a given dataset	A new criterion for selecting the best consensus data partition from a variety of consensus partitions
S19	Statistical Analysis and Data Mining	Cluster ensemble selection	Achieve statistically significant performance improvement over whole ensemble by explicitly considering quality and variety in ensemble selection	Replacement with other measure of quality and variety	Replacement with other quality measure and selection of a subset of a variety of solutions into a smaller cluster as well as better performance than using all available solutions
S20	International Joint Conferences on Artificial Intelligence	Adaptive Cluster Ensemble Selection	Better performance than the best team members to produce the ultimate solutions	compare to a state-of-the-art ensemble selection method	Introducing an adaptive cluster ensemble selection framework as a first stepe
S21	Pattern recognition	Hybrid clustering solution selection strategy	Provide good results and high performance using HCSS on most datasets	Use of hybrid clustering in large data sets in the fields of bioinformatics and data mining	use appropriate feature selection techniques to select clustering solutions.
S22	Intelligent Data Analysis	Cluster ensemble selection based on a new cluster stability measure	High performance of APMM standard compared to NMI proposed EEAC method	Investigating the effect of data sampling, variety and effect of noise and data loss	Propose a new clustering method based on subsets of all primary fake clustersI
S23	IEEE transactions on cybernetics	Transfer clustering ensemble selection	TCE-TCES can better balance quality and diversity, as well as produce more Suitable clustering results	Deploy TCE-TCES in a distributed environment to increase its performance and test it with different types of data sets, reviewing other hybrid strategies between transfer learning and CE	Propose a CES transfer algorithm that utilizes the relationship between quality and diversity in a source dataset
S24	Information Fusion	Moderate diversity for better cluster ensembles	The results suggest that selection by median diversity is no worse and in some cases is better than building and holding on to one ensemble	Find a combination of design discoveries, consensus functions and set size for a suitable data	Use the ARI to measure diversity in cluster groups and propose a diversity measure and provide accurate clustering in groups, also propose a procedure for constructing a cluster group.

ID	Journal	Title	Advantages	Disadvantages	Description
S25	Pattern recognition letters	Resampling-based selective clustering ensembles	The results obtained showed that the method of selective clustering sets based on re-sampling has a better solution compared to the methods of traditional clustering sets.	Most studies focus on the problem of creating a diverse group committee from a centralized clustering group and using similar methods or implementing a clustering algorithm.	Proposing a new method of clustering sets as a method of selective clustering sets based on re-sampling
S26	IEEE Access	Two-level-oriented selective clustering ensemble based on hybrid multi-modal metrics,	Selection of basic clustering partitions with variety and quality based on the proposed method and experimental analysis of the validity and stability of the proposed design	Most selective clustering algorithms evaluate diversity and quality with NMI and a combination of indicators, which are based on clustering labels without considering the data structure.	Proposing a new selective clustering group scheme, k-means combination and hierarchical clustering algorithm alternately with random design method in the production process of base clustering partitions to produce various base partitions
S27	Connection Science	A new method for weighted ensemble clustering and coupled ensemble selection	The high quality of the consensus obtained with this proposed method compared to the well-known clustering set algorithms in different benchmark datasets	Creating consensus based on surprising criteria at the cluster level based on the feasibility of selecting clusters, rather than clustering	Proposing a surprise measure at the cluster level to define clustering competence to reflect the level of agreement and disagreement between clusters
S28	In Australasian Database Conference	An Automatic Pruning Method Through Clustering Ensemble Selection	The results demonstrate that Auto-CES can effectively and efficiently prune the forest trees	Expand the algorithm in a large-scale environment including multi-cluster spark platforms.	Proposing a method for selecting Auto-CES for pruning random forest classifier (BC-RF) based on two main steps - clustering and selection
S29	International Journal of Autonomous and Adaptive Communications Systems	An efficient clustering ensemble selection algorithm.	Significantly improve clustering performance using the proposed algorithm	The existence of defects in the traditional selective clustering set and the lack of quality and accuracy and the fact that the selection of clustering partitions behave equally.	Proposing a new selective clustering group algorithm. Using the algorithm, first evaluate the validation of the clustering and select the best quality as the reference partition
S30	Journal of Intelligent and Fuzzy Systems,	Cluster ensemble selection using balanced normalized mutual information	High performance and better advanced cluster group methods with the proposed cluster set approach	Failure to consider a criterion for deciding on the participation of a cluster in a group	Development of a clustering set method based on cluster selection, inventing a standard called BNMI to test cluster stability to select a subset of the most stable cluster

ID	Journal	Title	Advantages	Disadvantages	Description
S31	Turkish Journal of Electrical Engineering and Computer Sciences	Clustering ensemble selection based on the extended Jaccard measure	The effectiveness and robustness of the proposed algorithm compared to the complete set	exploring the effects of noise and missing values of the data upon the EJ criterion and also on studying the application of the proposed method to different domains.	suggests a new hierarchical selection algorithm using a diversity/quality measure based on the Jaccard similarity measure
S32	International Conference on Computer Science and Engineering (UBMK)	Comparison of Different Clustering Ensembles by Solution Selection Strategy	The clustering set, especially the truncated BAGI, performs better than individual clustering methods by more accurately labeling data points, increasing robustness and effectiveness.	Combining multiple strategies in a single grouping model to better represent the available data, determining the number of automatically selected solutions is one of the problems of this method.	Design eight different groups of clustering using several clustering algorithms and compare in terms of accuracy with each other and evaluate the impact of these factors and propose a solution selection strategy based on accuracy
S33	International Conference on Social Computing and Social Media	Ensemble selection for community detection in complex networks	High performance	Planning in large-scale data sets to confirm preliminary results and compare with other group selection approaches based on tacit quality estimation	Proposing a diagram-based ensemble election approach and considering quality and diversity criteria and various quality criteria such as cluster-oriented quality and network-oriented quality functions
S34	arXiv preprint arXiv	ensemble selection using diversity and frequency	Improve clustering accuracy by evaluating natural data, especially considering the actual number of split clusters and the high performance of this method	Test this method using other diversity measures to find the optimal set size selected by the ESDF	Propose an efficient method for ensemble selection for a large ensemble and prioritize partitions in the set based on variability and frequency.
S35	In 2014 Seventh international conference on contemporary computing (IC3)	Leveraging frequency and diversity based ensemble selection to consensus clustering	Ensure the internal quality of clustering uniformly and without reduction with a greedy strategy for selecting clusters in a repetitive consensus generation technique and better clustering accuracy for the dataset	Testing the method using different criteria of diversity and importance of understanding the theoretical background of QPA with the criterion of general cluster quality for any desired cluster shape (Quality-based pair aggregation algorithm)	Investigate the need to select a subset of clusters to combine the best clusters of all existing clusters and overcome the impossible computational combination of partitions at the same time
S36	In 2016 IEEE International Conference on Automation Science and Engineering (CASE) (pp. 885-890). IEEE.	Model reduction method based on selective clustering ensemble algorithm and Theory of Constraints in semiconductor wafer fabrication	Guide construction managers to set the right timing rules based on a detailed model	Consider most of the dispatching rules	Propose a model reduction method based on clustering selection algorithm (SCEA) and constraint theory (TOC) to reduce computer runtime while maintaining the model's ability to correctly evaluate scheduling rules
S37	In International Workshop on Multiple Classifier Systems (pp. 179-189). Springer, Berlin, Heidelberg.	Selective clustering ensemble based on covariance	Improve clustering performance with the proposed algorithm	Further study of the case of selective clustering based on covariance and their use for practical applications, adding semi-regulatory information to this algorithm and achieving parallelization of this algorithm	Propose a method for measuring the diversity of basic clustering results and a covariance-based selective clustering set algorithm

ID	Journal	Title	Advantages	Disadvantages	Description
S38	In 2015 10th International Conference on Intelligent Systems and Knowledge Engineering (ISKE) (pp. 554-561). IEEE.	Selective Hierarchical Ensemble Modeling Approach and Its Application in Leaching Process	Using binary PSO optimization algorithm to find a group of MEHMs to reduce errors and increase variability	Improving the framework and methods with further studies in other industrial processes	Proposing a hybrid collection model (MEHM) based on the bagging algorithm. Proposing a new selective hierarchical set modeling approach to improve the accuracy and generalization of the set model and leaching model
S39	Fundamenta Informaticae, 176(1), 79-102	Social Network Optimization for Cluster Ensemble Selection	High performance of cluster group selection based on the proposed optimization compared to other complete set approaches	Improve modeling Optimization work to solve the optimal result for each IP model for large-scale datasets, solve the algorithm for the consensus function to automatically determine the appropriate number of clusters	Propose converting the similarity matrix to a modularity matrix and applying a new consensus function to optimize the modularity measurement
S40	In Journal of Physics: Conference Series (Vol. 1732, No. 1, p. 012074). IOP Publishing.	The Research on Clustering Ensembles Selection Algorithm based on Semi-supervised K-means Clustering.	Significantly improved performance compared to other clustering algorithms with the proposed algorithm	How to optimize the selected clustering algorithm and reduce the time complexity of the algorithm to have a better application algorithm	Proposing a new selective set algorithm based on semi-monitored K-means clustering. Check through a large number of tests for the validity of the proposed algorithm to deal with the clustering of high-dimensional data
ID	Journal	Title	Advantages	Disadvantages	Description
S41	2014 International Academic Conference of Postgraduates, NUAA	Wisdom of Crowds Cluster Ensemble Selection.	Checking the satisfaction of the relevant conditions and setting the main problems of the WOCCE algorithm with three threshold parameters on appropriate values	include decentralization criteria for generating primary results, independence criteria for the base algorithms, and diversity criteria for the ensemble members	Describing the WOC phenomenon to the problem of cluster set, introduction of social sciences, conditions of independence and decentralization in the field of cluster group research with WOC research.
S42	arXiv preprint arXiv: 2204.11062.	Selective clustering ensemble based on kappa and F-score	High efficiency and effectiveness of the proposed method	Overlap problems of clustering ensemble, community diagnosis ensemble	Using Kappa to select base partitions and F score for weight clusters as a new method for clusters and partitions leading to a new SCE method

Table 14: Comparison of Clustering Ensemble Selection methods

ID	year	DataSet	Diversity Approach	Diversity Measure	Consensus Function	consensus function result	Size of selected ensemble	Algorithm used for BC
S1	2015	Soybean, Ecoli, Breast tissue, Iris, Wine, Glass, Breast cancer, Satimage	Non Pairwise-Hybrid	NMI	CSPA, HGPA	for combining the full ensemble members and combining different subsets of full ensemble members.	Automatic	K-means
S2	2017	Waveform, segmentation, Statlog, Thyroid, Soybean, Iris, Hearts, Wine, SPECTheart, Glass, Lung	Pairwise	NMI	spectral clustering, HGPA, CSPA	to combine the solutions to produce a final consensus clustering and show the results of the CSPA approach	Automatic	k-means
S3	2019	Breast-cancer, Iris, Bupa, Saheart, Ionosphere, Glass, Halfring, Galaxy, Yeast, Wine	Non-pairwise	NMI	CSPA, HGPA, MCLA, Average link	for deriving the final clusters from co-association matrix	fixed	k-means
S4	2013	Iris, Wine, Breast, Chart, Yeast, Articles, cbrilpirivson	Pairwise	multiple criteria	CSPA, HGPA, MCLA, Average Linkage	seems to be favored by the quality and not by the cardinality or diversity of the selected	fixed	k-means
S5	2011	Iris, Wine, Segmentation, Heart, Lung, wdbc, Sat. image, ionosphere	Non-pairwise	NMI, ARI	CSPA, MCLA	consensus function is needed to combine clustering and produce a final partition	fixed	Spectral Clustering, Nyström approximation, random scaling parameter and random initialization of k-means
S6	2020	Breast Cancer, Ecoli, Glass Identification, Iris, Lung Cancer, Seeds, Soybean (Small), Statlog (Heart), Wine, Yeast	Pairwise	NMI, ARI, JI	Normalized Cut Algorithm	to get the final result and get performance clustering robustness.	fixed	k-means, Spectral Clustering
S7	2017	Zoo, promoter, hayes, dermatology, vote, balance, breastcancer, krvskp, mushroom, nursery, basehock, pcmac	Non-pairwise	CA NMI, ARI	SL, CL, CSPA, HGPA, MCLA	Using the CSPA consensus function, the SIVID algorithm obtains the best performance on eight of the twelve data sets	fixed	k-modes clustering algorithm

S8	2019	S1, Jain, Flame, Pathbased, Aggregation, D31, Iris, Heart, Wine, Protein localization sites, Australian credit approval, Waveform	pairwise	CA, NMI, ARI, SC, CHI	k-DPP	to select k base clusterings from M candidates for clustering ensemble	fixed	K-means
S9	2018	Iris, Wine, Seeds, Glass, Protein Localization Sites, Ecoli, LIBRAS Movement Database, User Knowledge Modeling, Vote, Wisconsin Diagnostic Breast Cancer, Synthetic Control Chart Time Series, Student, Australian Credit Approval, Cardiotocography, Wave form Database Generator, Parkinsons Telemonitoring, Statlog Landsat Satellite, Tr12, Tr11, Tr45, Tr41, Tr31, Wap, Hitech, Fbis	pairwise	NMI, AC, ARI	average-link hierarchical clustering algorithm based on the CO-matrix	Useful in reflecting the performance of selected partitions	fixed	k-means algorithm
S10	2012	Iris, Segmentation, Lung, WDBC, Sat.image	pairwise	NMI, ARI	CSPA, HPGA and MCLA,	The HPGA approach works poorly in consensus clustering. CSPA and MCLA are considered, but only MCLA is used for complexity	fixed	Spectral Clustering Nyström approximation and random initialization of k-means
S11	2016	Half Ring, Iris, Balance Scale, Breast Cancer, Bupa, Galaxy, Glass, Ionosphere, SA Heart, Wine, Yeast, Pendigits, Statlog, Optdigits, Arcene, CNAE-9, Sonar	Non-pairwise	NMI. APMM	WOCCE, APMM, MAX, and MCLA	show the effect of the aggregation method on improving accuracy in the final results, WOCCE and the proposed algorithm have generated better results in comparison with other basic and ensemble algorithms.	fixed	K-means, multiple algorithm
S12	2015	Zachary, US Politics, Dolphins	pairwise	NMI, ARI, VI	CSPA	The approach is based on constructing a consensus graph out of the set of partitions to be combined	fixed	relative neighborhood graphs (RNG), neighborhood graph, k-nearest neighbor graph

S13	2012	IRIS, WDBC, Soybean	Non-pairwise	NMI	CSPA	to produce the final result	fixed	k-means
S14	2014	Problem 1 (blobs), Problem 2 (moon)	pairwise	Davies-Bouldin Index, Beni-Xie Index, Eigengap Index	consensus clustering is given by the aggregate membership matrix(KM, FCM, AGPCM, SC(fix), SC(var1), SC(var2).)	Although the overall purity of the ensembles is slightly smaller than that of the best performing clusterings, the method clearly points out the most suitable paradigm in each case.	fixed	k-means, fuzzy c-means, Asymmetric Graded Possibilistic c-Means, Spectral Methods
S15	2019	Landsat, Dna, Wpb, Vehicle, Vote, Heart, Breast, Msplice	pairwise	NMI	CSP, HGP, MCL, DP	DP algorithm is logical and can further improve clustering performance	fixed	single spectral clustering
S16	2011	Iris, Wine, Segmentation, Heart, Lung, WDBC Sat.image, Ionosphere, Vehicle, Sonar	pairwise	ARI	CSPA, HPGA, and MCLA	only MCLA and HPGA are used for complexity	fixed	spectral clustering by projecting the original sample space into a low dimension space, Nyström approximation, the random scaling parameter, and the random initialization of k-means
S17	2021	Adult, Australian, cancer, census, coil2000, column, credit, crowd, eeg, fars, flare, FPS-5, german, letter, magic, market, nursery, optdigits, ring, sick, spambase, thyroid, twonorm, waveform, wine, Number of bests	pairwise	convenience	Algorithm (ES-JSS)(base learners)	ES-JSS achieved the best performance on %64 data sets	fixed	spectral clustering,, structural sparsity

S18	2009	Bars Breast C. Cigar Half Rings Iris Log Yeast Std Yeast Optdigits Spiral	Non pairwise	ANM I	Single-Link (SL), Average-Link (AL), Complete-Link (CL), K-means (KM), CLARANS (CLR), Chameleon (CHM), CLIQUE, CURE, DBSCAN and STING	applied the EAC, SWEACS and JWEACS approaches using the KM, SL, AL andWard-Link (WR) [23] clustering algorithms to produce the consensus partitions.	fixed	Single-Link and K-means
S19	2008	CBIR, CHART, EOS, ISOLET6, SEGMENTATION, WINE	Pairwise-Non Pairwise	NMI	CSPA	to obtain a consensus clustering solution, whose NMI value is then computed using the class label information	Fixed	K-means
S20	2009	Iris, Soybean, Wine, Thyroid	Non Pairwise	NMI	HAC-AL	robust to the choice of the consensus function	Fixed	K-means, Maximal similar features
S21	2014	Derma, Breast, Heart, Soybean, Image, Ecoli, Seeds, Lymphoma, SRBCT, Gliomas, ET-CNS, M-tissue	Non-pairwise	NMI, dominant ratio, Squared-Error distortion, Disassociation	Normalized cut Algorithm	Ncut is able to provide more accurate and stable results, and is more suitable to serve as the consensus function when compared with KM, SC and SOM	Fixed	K-means, spectral clustering
S22	2014	Breast-Cancer, Iris, Bupa, SA-Heart, Ionosphere, Glass, HalfRing, Galaxy, Yeast, Wine	pairwise	APMM	Average link	the most effective consensus function results from the average-linkage hierarchical clustering algorithm	Fixed	K-means
S23	2018	OVERVIEW OF THE 20NG DATASET	Pairwise, Hybrid	NMI, ARI	Normalized cut Algorithm	can be applied to partition	Fixed	K-means, spectral clustering

S24	2006	Four-gauss, Easy-doughnut, Difficultdoughnut, Glass, Wine	Pairwise-Non Pairwise	ARI	K-means	The consensus function was k-means clustering using the consensus matrix as the input data. This choice was based on a small pilot set of experiments which showed this consensus function to be superior to the one used before for the current set-up	Fixed	K-means, hierarchical clustering algorithm
S25	2009	IRIS, WINE, HEART, LUNG, WDBC, VEHICLE, SEGMENTATION, SAT. IMAGE	Pairwise-Non Pairwise	Rand Index method	CSPA, HGPA, MCLA	one with the maximum average normalized mutual information is returned as the final clustering result, can achieve better solutions	Fixed	K-means
S26	2018	Iris, wine, breast cancer, pima indian woman diabetes(pima 1), pima Indians diabetes(pima2)	Pairwise-Non Pairwise	NMI, Tanimoto coefficient, Silhoutte coefficient, CH	selective clustering ensemble algorithm MMSCE based on multi-modal metrics	Basic clustering partitions with variety and high quality as well as a reliable source for clustering selection	Fixed	K-means, hierarchical clustering algorithm with random projection
S27	2021	Iris, Wine, Glass, IS, Ecoli, SPF, Yeast, Avila, LR	pairwise	NMI	WHAC, CES, LWEA, EAC, WEAC-AL, GP-MGLA, PTA-AL, PTA-CL, PTA-SL, PTGP, CSPA, HGPA, MCLA, IVC, IPVC, IPC, CAS	can be concluded that CES(coupled ensemble selection) outperforms other methods in most of the data sets and as an efficient ensemble selection technique	Fixed	K-means

S28	2018	Soybean, Breast cancer, Wilt, Sonar	pairwise	F-measure	CLUB-DRF applies the K-MODES clustering model to group the trees	Collect the most accurate trees based on the area under the curve	Fixed	Breiman as CART-based RF (BC -RF)
S29	2015	GLASS, IRIS, WDBC, Soybean, Heart, WINE	pairwise	ARI	CSPA	to fuse to get the final result	Fixed	k-means
S30	2020	Wine, Breast-Cancer, Bupa, Ionosphere, Iris, Glass, WDBC, Yeast, Galaxy, SAHeart, Image, Lymphoma, OQ, Pima, Sonar, MNIST 1vs2, MNIST, 2-Spiral, Aggregation, Flame, 3-Spiral, Open Flame, Halfring	pairwise	NMI	CSPA, HGPA and MCLA, FCM, hierarchical, E-EAC	the best option for consensus function is to apply the algorithm named average-linkage on E-EAC-based co-association extracted by ItoU equation.	Fixed	kmeans
S31	2021	Jain, Path bass, Aggregation, Soybean(small), Breast-tissue, Iris, Wine, Seeds, Glass, Ecoli, Breast-cancer, Yeast, Segmentation, Satimage	pair-wise, hybrid	EJ	CSPA, HGPA and MCLA,	used to obtain the consensus solution and cluster ensemble selection results with a hierarchical method	Fixed	k-means
S32	2018	Blogger, car, dermatology, ecoil, haberman, heart-statlog, hepatitis, iris, lymphography, segment, seismic-bumps, sick, wine, zoo	pairwise	Using accuracy criteria	Single, voted (seeds), pruned, bagi, pruned (bagi), Bag2, rs, pruned (rs)	Clustering ensemble, especially pruned BAGI, outperform single clustering methods by labelling the data points more accurately while increasing the robustness and effectiveness	Fixed	k-means, EM (expectation maximization), hierarchical, canopy, farthest first,
S33	2015	Zachary, US Politics, Dolphins	pairwise	ARI, NMI,	CSPA	compute a consensus partition applying a CSPA ensemble clustering approach on the whole set of obtained partitions and set of partitions selected	fixed	Graph-based cluster ensemble selection algorithm
S34	2015	Chart, Segmentation, Ecoli, Yeast, Iris, Glass, Wine, Vehicle	pairwise	ARI	CSPA, HGPA,	shows datasets, CSPA and HGPA produce better consensus when ESDF is used as the ensemble selection procedure rather than CAS	fixed	k-means

S35	2014	Chart, Segmentation, Ecoli, Yeast, Iris, Glass, Wine, Vehicle	pairwise	ARI	CSPA/HGPA	CSPA and HGPA produce better consensus when ESDF is used as the ensemble selection procedure instead of CAS	fixed	k-means
S36	2016	24 workstations and 72 machines	pairwise	MID	selective clustering ensemble algorithm (SCEA), Theory of Constraints (TOC)	Measure the importance of machinery comprehensively	fixed	k-means
S37	2013	Iris, Wine, Zoo, Glass, Ionosphere, Sonar, Balance scale, Pima, Spectheart, Hepatitis, Bupa, Habermans survival, Wdbc, Statlog, Vehicle, Breast-cancer-Wisconsin, Car, Credit-g, Vowel, Lymphography	pairwise	covariance	CSPA(ALL, RSE, CSEV)	ALL is directly ensemble, RSE is selective ensemble based on random, and CSEV is average value of selective ensemble based on covariance,, We can obtain two conclusions based on above results. Firstly, the clustering ensemble result is better than base clustering. Secondly, the CSEV is better than base clustering, ALL, and RSE,	fixed	K-Means, AP, and FCM
S38	2015	dataset with thirty-six groups of data can be obtained and each group has twelve samples. The dataset is divided into two sets averagely	pairwise	root mean squared error (RMSE) and maximal absolute error (MAXE)	bagging ensemble (MEHM), particle swarm optimization (PSO) algorithm	the NSMEHM does consistently improve the predicted precision versus MM, SVM and MEHM for leaching process	fixed	SVM algorithm and base vector (BV) bootstrap sampling algorithm

S39	2020	BreastTissue, Iris, Wine, Glass Identification, Haberman's Survival, Vertebral Column 3C, Ecoli, Liver Disorders Bupa, Digits, Yeast, Half Rings.	pairwise	AAPMM	EEAC(Single, Average, Complete), Modularity (Sum Link)	the proposed sum linkage algorithm as the modularity based consensus function of cluster ensemble selection is definitely the best option to cluster an input data.,,	fixed	K-Means,
S40	2021	Wine, Waveform(version1), TSE, Libras Movement, WLEMIMU	pairwise	NMI	clustering ensemble algorithm based on semi-supervised K-means clustering(skc), SKCSE(without reference partition),	projects the high-dimensional space data to the low-dimensional space, improve the accuracy of the initial cluster member.	fixed	k-means algorithm and the Semi-supervised k-means clustering algorithm
S41	2014	Half Ring, Iris, Balance Scale, Breast Cancer, Bupa, Galaxy, Glass, Ionosphere, SA Heart, Wine, Yeast, Pendigits, Statlog, Optdigits	pairwise	NMI	MCLA, MAX and WOCCE	MCLA, MAX and WOCCE have generated better results in comparison with CSPA and HGPA	fixed	using different algorithms and changing the number of partitions
S42	2022	Iris, Wine, Seeds, Glass, Protein Localization Sites, Ecoli, LIBRAS Movement Database, User Knowledge Modeling, Vote, Wisconsin Diagnostic Breast Cancer, Synthetic Control Chart Time Series, Australian Credit Approval, Cardioto cography, Wave, form, Database Generator, Parkinsons Telemonitoring, Statlog Landsat Satellite, Tr12, Tr11, Tr45, Tr41, Tr31, Wap, Hitech, Fbis	pairwise	use kappa and F-score as evaluation metrics, instead of NMI	Hierarch ical agglomerative clustering with average linkage (HAC-AL)	to generate the final partition	Fixed	K-means

Figure 6: Diagram based on the year number of studies

4.2 RQ2: What is the diversity?

In general, clustering methods are divided into partition categories and hierarchical methods. A single clustering in partition methods returns the final clusters, and in hierarchical methods, nested clusters return the dataset obtained from cumulative algorithms and partitioning algorithms. The point algorithm considers each point (pattern) as a cluster and identifies and merges the nearest cluster to create the next cluster. Dividing algorithms select the clusters produced in each step and divide them into two smaller clusters. There are some basic clustering algorithms; a simple algorithm, called the k-means algorithm, has been used by many researchers. The k-means clustering algorithm is applied as a partition classification, only in numerical data sets[36]. In the k-means algorithm, K clusters are developed so that the points of the cluster itself are closer to the center of their corresponding cluster than the center of the other clusters. By selecting the K points that are the center of the cluster, the algorithm process begins. By the selection of the points, these points, which are assigned to the nearest center, create clusters. The average points are then measured as centers, which are the average vectors. Eventually, this process will produce a new cluster by the new center[26]. The algorithm will run until the centers change. The steps of the k-means algorithm (K-means algorithm to find k clusters) are shown in the following algorithm:

1. Select k points as the centers of the clusters
2. Assign all points to closer centers and create k clusters
3. Redesign the centers of the clusters
4. Ensure that the central points of the clusters do not change by repeating steps 2 and 3.

In addition, hierarchical algorithms include Single link [46], Average link [40], and Complete link [31]. If two partitions are different, the labels of one partition are not the same as the labels of the other partitions. Normalized Mutual Information (NMI)[47] and Modified Rand Index (ARI) [25] are used for partition quality and diversity measurement. The ARI and NMI quality criteria are obtained by the following method:

Normalized Mutual Information (NMI): The Normalized Mutual Information proposed by [47] can be defined as follows:

$$NMI(\pi_a, \pi_b) = \frac{-2\sum_{i=1}^{k_a}\sum_{j=1}^{k_b} n_{ij}\log(\frac{n.n_{ij}}{n_{ia}.n_{bj}})}{\sum_{i=1}^{k_a} n_{ia}\log(\frac{n_{ia}}{n}) + \sum_{j=1}^{k_b} n_{bj}\log(\frac{n_{bj}}{n})} \tag{4.1}$$

Adjusted Rand Index (ARI): The Adjusted Rand Index[25] is defined as follows:

$$ARI(\pi_a, \pi_b) = \frac{\sum_{i=1}^{k_a} \sum_{j=1}^{k_b} \binom{n_{ij}}{2} - t_3}{\frac{1}{2}(t_1 + t_2) - t_3} \tag{4.2}$$

where,

$$t_1 = \sum_{i=1}^{k_a} \binom{n_{ia}}{2}, \qquad t_2 = \sum_{j=1}^{k_b} \binom{n_{bj}}{2}, \qquad t_3 = \frac{2t_1t_2}{n(n-1)} \tag{4.3}$$

Diversity measures could be separated into pair-wise,non-pair wise and hybrid.[30] The selected articles used three methods of diversity approach, which are 58% pair-wise, 34% non pair-wise, and 8% Hybrid. Table 10 and Figure 7 show the number of studies on the methods and diversity approach, respectively. It can be seen that three methods have been studied, mostly in pair-wise methods with 58%, then non pair-wise with 34%, and finally Hybrid with 8%.

Table 15: Diversity Approach in the clustering ensemble selection

NO.	Method	%	Studies ID
1	Pairwise	58	S2, S4, S6, S8, S9, S10, S12, S14, S15, S16, S17, S19, S22, S23, S24, S25, S26, S27, S28, S29, S30, S31
2	Non-pairwise	34	S1, S3, S5, S7, S11, S20, S13, S18, S19, S21, S24, S25, S26
3	Hybrid	8	S1, S23, S31

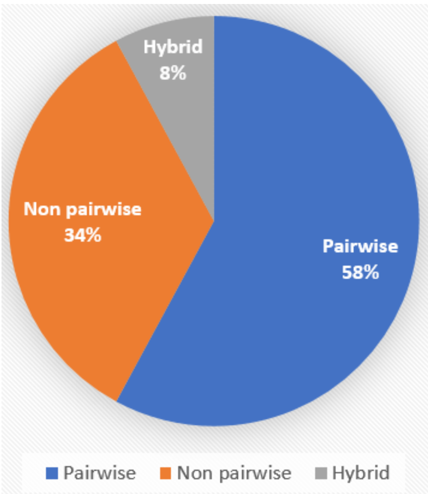

Figure 7: Diversity Approach

4.3 RQ3: How base clusterings are generated in different methods?

For diversity generation, there are different methods of base clustering, which are fully described in 3.4. According to the studies performed on the articles listed in Table 11, 20 articles from the method number one, 13 articles from the method number 2, 7 articles from the method number 3, and 4 articles from the method number 5 have been used for diversity generation (base clustering). According to the table presented below, the most articles (20%) were of the method number 1 and the least articles (4%) were of the method number 5 (see Table 11 and Figure 8).

Table 16: Generate Steps For Basic Clustering in the clustering ensemble selection

NO.	Generate Diversity	%	Studies ID
1	a	45	S23, S1, S8, S7, S3, S9, S2, S4, S13, S42, S22, S27, S35, S34, S33, S31, S30, S29, S39, S19
2	b	30	S10, S15, S6, S21, S5, S18, S24, S12, S20, S14, S41, S38, S37
3	c	16	S19, S25, S11, S17, S36, S32, S28
4	e	9	S26, S16, S40, S19

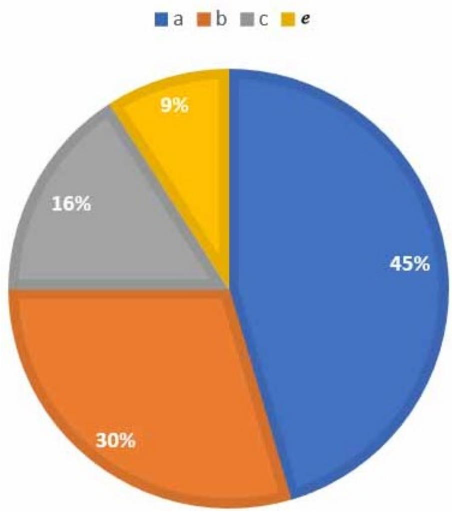

Figure 8: Generate Steps For Basic Clustering

4.4 RQ4: Which measures are worked in CES?

Some criteria are useful for evaluating the quality of data partitions, e.g., quantitative criteria. Most of the cluster validity criteria could be separated into two groups of internal and external criteria. Internal criteria examine the structure of data using a clustering algorithm considering a criterion defined between data, as well as clustering without resorting to the reference partition. On the other hand, external criteria measure the difference between a structure based on the class label and the structure defined by a cluster. Here are some commonly used measure: And the following table lists the number of measures used.

Internal quality measures:

Davies-Bouldin Index (DBI): The Davies-Bouldin Index [10] is defined as follows:

$$DBI_k = \frac{1}{k}\sum_{h=1}^{k} F_{C_h} \tag{4.4}$$

where,

$$F_{C_h} = \max_{C_j \neq C_h} F_{C_h C_j}, \qquad F_{C_h C_j} = \frac{f_1(C_h) + f_1(C_j)}{f_2(C_h, C_j)} \tag{4.5}$$

Silhouette Index (SI): The Silhouette Index [29] is defined as follows:

$$SI(k) = \frac{1}{k}\sum_{h=1}^{k} SI_h \tag{4.6}$$

where,

$$SI_h = \frac{1}{|C_h|}\sum_{i=1}^{|C_h|}\left[\frac{b_i^h - a_i^h}{max\{a_i^h, b_i^h\}}\right] \tag{4.7}$$

$$a_i^h = \frac{1}{|C_h| - 1}\sum_{l=1, l\neq i}^{|C_h|} d(x_i^h, x_l^h), \qquad b_i^h = \min_{j\in\{1,\cdots,k\}, j\neq h}\left\{\frac{1}{|C_j|}\sum_{l=1}^{|C_j|} d(x_i^h, x_l^j)\right\} \tag{4.8}$$

External quality measures:

Disagreement and Agreement Index (DAI): The Disagreement and Agreement Index was proposed by [57] as an external measure. DAI is defined as:

$$DAI(k) = \frac{1}{L}\sum_{i=1}^{L} \tau_k(\pi^*, \pi_i) \tag{4.9}$$

where,

$$\tau_k(\pi^*, \pi_l) = \frac{\sum_{i<j} 1\{m^*_{ij} \neq m^l_{ij}\}}{\sum_{i<j} 1\{m^*_{ij} = m^l_{ij}\}}, i = \{1, \cdots, k^*\}, j = \{1, \cdots, k_l\} \tag{4.10}$$

$$m^l_{ij} = \begin{cases} 1, & x_i \text{ and } x_j \text{ are in clustering } \pi_l; \\ 0, & \text{else.} \end{cases} \quad i, j = \{1, \cdots, k_l\} \tag{4.11}$$

F-measure (FM): The F-measure (F-score) [33] is defined as follows:

$$FM(\pi_a, \pi_b) = max \sum_{i=1}^{k_a} \frac{2 \times n_{ia} \times \left(\frac{n_{ij}}{n_{ia}} + \frac{n_{ij}}{n_{jb}}\right)}{n \times \left(\frac{n_{ij}}{n_{ia}} + \frac{n_{ij}}{n_{jb}}\right)} \tag{4.12}$$

Selection of clusterings:

Recently, a little research has concentrated heuristically on how to select subset of ensemble members considering quality and diversity [35, 3].

Selective clustering ensemble based on covariance (SCEBC): A diversity measure was introduced by [35] considering the covariance. **CES based on APMM criterion**: The authors in [3] introduced a novel criterion, called Alizadeh-Parvin-Moshki-Minaei (APMM) as well as an innovative method called Extended Evidence Accumulation Clustering (EEAC). which can be computed by means of Eq. (4.13).

$$APMM(C^a_i, P^{b^*}) = \frac{-2\; n^a_i \; log\left(\frac{n}{n^a_i}\right)}{n^a_i \; log\left(\frac{n^a_i}{n}\right) + \sum_{j=1}^{k_{b^*}} n^{b^*}_j \; log\left(\frac{n^{b^*}_j}{n}\right)} \tag{4.13}$$

Each entry of the co-association matrix in this method is computed as follows:

$$C(i, j) = \frac{n_{ij}}{max(n_i, n_j)} \tag{4.14}$$

The types of internal and external majors used in the articles are listed according to Table 12. According to the reviews conducted on the articles, the NMI measure has been used more.

4.5 RQ5: Which journal have paid more attention to CES ?

All the resources, various publication channels, and the number of papers per publication source are presented in Table 13. Three publication channels were determined: journal, conference, and workshop. Among the 42 selected studies,26 papers (62%) had been published in journals, 14 papers (33%) had been presented at conferences, and 2 papers (5%) came from a workshop. Table 13 demonstrates the distribution of the selected studies in terms of the publication sources, and Figure 9 shows the publication venue.

5 Conclusion and future work

This systematic mapping study (SMS) analyzed and synthesized articles related to clustering ensemble selection. This is an effective technique for improving the quality of clustering solutions. A total of 42 articles were published by Hadjitodorov from 2006 to August 2022, based on the year of publication. Basic clustering was used to generate diversity and the criteria applied to composite clustering. the most of the articles were published in 2015 and the

Table 17: Diversity Measure in the clustering ensemble selection

NO.	Diversity Measure	Studies ID	NO.	Diversity Measure	Studies ID
1	NMI	S1, S2, S3, S5, S6, S7, S8, S9, S10, S11, S12, S13, S15, S19, S20, S21, S23, S26, S27, S30, S33, S40, S41	17	Dominant raito	S21
2	ARI	S5, S6, S7, S8, S9, S10, S12, S16, S23, S24, S29, S33, S34, S35	18	Squared Error Distortion	S21
3	Multiple criteria	S4	19	Disassociation	S21
4	JI	S6	20	RI method	S25
5	CA	S7, S8	21	Tanimoto coefficient	S26
6	SC	S8	22	Silhoutte coefficient	S26
7	CHI	S8	23	CH	S26
8	AC	S9	24	F-measure	S28, S42
9	APMM	S11, S22	25	Ej	S31
10	VI	S12	26	Accuracy criteria	S32
11	Davies Bouldin Index	S14	27	MID	S36
12	Beni Xie Index	S14	28	Covariance	S37
13	Eigengap Index	S14	29	RMSE	S38
14	covariance	S37	30	MAXE	S38
15	F-measure	S28, S42	31	AAPMM	S39
16	ANMI	S18	32	Kappa	S42

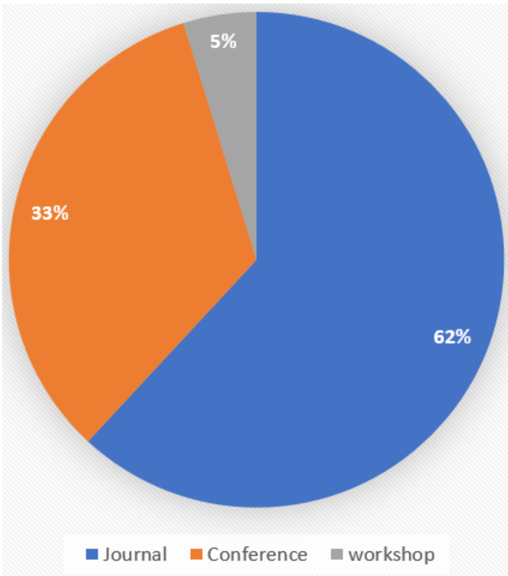

Figure 9: Publication Venue

smallest number of them in 2006 and 2008. The pair-wise diversity with 58% was a diversity method that was most frequently used in clustering ensemble selection. In addition, most of the articles have used the NMI measure to evaluate the cluster quality,and the method of valuing the initial parameter has been more-commonly used for the generation of diversity. According to the results of this research, the trade-off between diversity and quality (considering both at the same time) can be studied and evaluated in the future. Moreover, clustering ensemble selection has not been done on text yet, which is a gap recommended to be filled by future research.

Table 18: Publication venues

P.Ch*	Publication venue (Number of studies)
Journal	Engineering Applications of Artificial Intelligence(2)
	Neurocomputing(1)
	Artificial Intelligence Review(1)
	Data Mining and Knowledge Discovery(1)
	Pattern Recognition Letters(2)
	Pattern Recognition(3)
	Soft Computing(1)
	ACM Transactions on Knowledge Discovery from Data(1)
	In *Recent Advances of Neural Network Models and Applications,* Springer, Cham(1)
	Wuhan University Journal of Natural Sciences(1)
	Statistical Analysis and Data Mining(1)
	Intelligent Data Analysis(1)
	IEEE transactions on cybernetics(1)
	Information Fusion(1)
	IEEE Access(1)
	Connection Science(1)
	International Journal of Autonomous and Adaptive Communications Systems(1)
	Journal of Intelligent & Fuzzy Systems(1)
	Turkish Journal of Electrical Engineering & Computer Sciences(1)
	arXiv preprint arXiv(2)
	Fundamenta Informaticae(1)
Conference	*International Conference on Neural Information Processing,* Springer(1)
	International Conference on Fuzzy Systems and Knowledge Discovery, IEEE(1)
	In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*(1)
	In *2012 IEEE Ninth International Conference on e-Business Engineering* , IEEE(1)
	In *International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management,* Springer(1)
	International Joint Conferences on Artificial Intelligence(1)
	In *Australasian Database Conference*(1)
	International Conference on Computer Science and Engineering(1)
	International Conference on Social Computing and Social Media(1)
	In *2014 Seventh international conference on contemporary computing* (1)
	In *2016 IEEE International Conference on Automation Science and Engineering,IEEE*(1)
	In *2015 10th International Conference on Intelligent Systems and Knowledge Engineering* IEEE(1)
	In *Journal of Physics: Conference Series,* IOP Publishing(1)
	2014 International Academic Conference of Postgraduates, NUAA(1)
workshop	In *International Workshop on Multiple Classifier Systems,*Springer, Berlin, Heidelberg. (1)
	In *International Workshop on Multiple Classifier Systems* (pp. 179-189). Springer, Berlin, Heidelberg. (1)

References

[1] D.D. Abdala, P. Wattuya, and X. Jiang, *Ensemble clustering via random walker consensus strategy*, 20th Int. Conf. Pattern Recogn., IEEE, 2010, pp. 1433–1436.

[2] M.T. AL-Sharuee, F. Liu, and M. Pratama, *Sentiment analysis: An automatic contextual analysis and ensemble clustering approach and comparison*, Data Knowledge Engin. **115** (2018), 194–213.

[3] H. Alizadeh, B. Minaei-Bidgoli, and H. Parvin, *Cluster ensemble selection based on a new cluster stability measure*,

Intel. Data Anal. **18** (2014), no. 3, 389–408.

[4] ______, *To improve the quality of cluster ensembles by selecting a subset of base clusters*, J. Experimen. Theor. Artific. Intel. **26** (2014), no. 1, 127–150.

[5] J. Azimi and X. Fern, *Adaptive cluster ensemble selection*, Twenty-First Int. Joint Conf. Artific.l Intel., 2009.

[6] L. Bai, J. Liang, H. Du, and Y. Guo, *An information-theoretical framework for cluster ensemble*, IEEE Trans. Knowledge Data Engin. **31** (2018), no. 8, 1464–1477.

[7] V. Berikov, *Weighted ensemble of algorithms for complex data clustering*, Pattern Recogn. Lett. **38** (2014), 99–106.

[8] I. Bifulco, C. Fedullo, F. Napolitano, G. Raiconi, and R. Tagliaferri, *Robust clustering by aggregation and intersection methods*, Int. Conf. Knowledge-Based Intel. Inf. Engin. Syst., Springer, 2008, pp. 732–739.

[9] T. Boongoen and N. Iam-On, *Cluster ensembles: A survey of approaches with recent extensions and applications*, Comput. Sci. Rev. **28** (2018), 1–25.

[10] D.L. Davies and D.W. Bouldin, *A cluster separation measure*, IEEE Trans. Pattern Anal. Machine Intel. (1979), no. 2, 224–227.

[11] U.M. Fayyad, C. Reina, and P.S. Bradley, *Initialization of iterative refinement clustering algorithms.*, KDD, 1998, pp. 194–198.

[12] X.Z. Fern and C.E. Brodley, *Random projection for high dimensional data clustering: A cluster ensemble approach*, Proc. 20th Int. Conf. Machine Learn. (ICML-03), 2003, pp. 186–193.

[13] ______, *Solving cluster ensemble problems by bipartite graph partitioning*, Proc. Twenty-First Int. Conf. Machine Learn., 2004, p. 36.

[14] X.Z. Fern and W. Lin, *Cluster ensemble selection*, Statist. Anal. Data Min.: ASA Data Sci. J. **1** (2008), no. 3, 128–141.

[15] A.L.N. Fred and A.K. Jain, *Combining multiple clusterings using evidence accumulation*, IEEE Trans. Pattern Anal. Machine Intel. **27** (2005), no. 6, 835–850.

[16] A. Ghosh, J. Acharya, *Cluster ensembles*, Wiley Interdiscip. Rev.: Data Min. Knowledge Disc. **1** (2011), no. 4, 305–315.

[17] J. Ghosh, A. Strehl, and S. Merugu, *A consensus framework for integrating distributed clusterings under limited knowledge sharing*, Proc. NSF Workshop on Next Generation Data Mining, Citeseer, 2002, pp. 99–108.

[18] S.T. Hadjitodorov, L.I. Kuncheva, and L.P. Todorova, *Moderate diversity for better cluster ensembles*, Information Fusion **7** (2006), no. 3, 264–275.

[19] Y. Hong, S. Kwong, Y. Chang, and Q. Ren, *Unsupervised feature selection using clustering ensembles and population based incremental learning algorithm*, Pattern Recogn. **41** (2008), no. 9, 2742–2756.

[20] X. Hu, *Integration of cluster ensemble and text summarization for gene expression analysis*, Proc. Fourth IEEE Symp. Bioinf. Bioengin., IEEE, 2004, pp. 251–258.

[21] X. Hu and I. Yoo, *Cluster ensemble and its applications in gene expression analysis*, Proc. Second Conf. Asia-Pacific Bioinf. Volume 29, 2004, pp. 297–302.

[22] D. Huang, J. Lai, and C.-D. Wang, *Ensemble clustering using factor graph*, Pattern Recogn. **50** (2016), 131–142.

[23] D. Huang, C.-D. Wang, and J.-H. Lai, *Locally weighted ensemble clustering*, IEEE Trans. Cybernet. **48** (2017), no. 5, 1460–1473.

[24] ______, *Lwmc: A locally weighted meta-clustering algorithm for ensemble clustering*, Int. Conf. Neural Inf. Process., Springer, 2017, pp. 167–176.

[25] L. Hubert and P. Arabie, *Comparing clusterings*, J. Classific. **2** (1985), 193–218.

[26] A.K. Jain and R.C. Dubes, *Algorithms for clustering data*, Prentice-Hall, Inc., 1988.

[27] A.K. Jain, M.N. Murty, and P.J. Flynn, *Data clustering: A review*, ACM Comput. Surv. (CSUR) **31** (1999), no. 3, 264–323.

[28] V. Kandylas, S. Upham, and L.H. Ungar, *Finding cohesive clusters for analyzing knowledge communities*, Knowledge Inf. Syst. **17** (2008), no. 3, 335–354.

[29] L. Kaufman and P.J. Rousseeuw, *Finding groups in data: An introduction to cluster analysis*, John Wiley & Sons, 2009.

[30] H. Khalili, M. Rabbani, and E. Akbari, *Clustering ensemble selection based on the extended Jaccard measure*, Turk. J. Electric. Engin. Comput. Sci. **29** (2021), no. 4, 2215–2231.

[31] B. King, *Step-wise clustering procedures*, J. Amer. Statist. Assoc. **62** (1967), no. 317, 86–101.

[32] H.W. Kuhn, *The Hungarian method for the assignment problem*, Naval Res. Logistics Quart. **2** (1955), no. 1-2, 83–97.

[33] B. Larsen and C. Aone, *Fast and effective text mining using linear-time document clustering*, Proc. Fifth ACM SIGKDD Int. Conf. Knowledge Disc. Data Min., 1999, pp. 16–22.

[34] T. Li and C. Ding, *Weighted consensus clustering*, Proc. SIAM Int. Conf. Data Min., SIAM, 2008, pp. 798–809.

[35] X. Lu, Y. Yang, and H. Wang, *Selective clustering ensemble based on covariance*, Int. Workshop Multiple Classifier Syst., Springer, 2013, pp. 179–189.

[36] J. MacQueen, *Classification and analysis of multivariate observations*, 5th Berkeley Symp. Math. Statist. Probability, 1967, pp. 281–297.

[37] S. Mimaroglu and M. Yagci, *CLICOM: cliques for combining multiple clusterings*, Expert Syst. Appl. **39** (2012), no. 2, 1889–1901.

[38] B. Minaei-Bidgoli, A. Topchy, and W.F. Punch, *Ensembles of partitions via data resampling*, Int. Conf. Inf. Technol.: Cod. Comput., 2004. Proc. ITCC 2004., vol. 2, IEEE, 2004, pp. 188–192.

[39] S. Nejatian, H. Parvin, and E. Faraji, *Using sub-sampling and ensemble clustering techniques to improve performance of imbalanced classification*, Neurocomput. **276** (2018), 55–66.

[40] C.F. Olson, *Parallel algorithms for hierarchical clustering*, Parallel Comput. **21** (1995), no. 8, 1313–1325.

[41] Y. Ren, C. Domeniconi, G. Zhang, and G. Yu, *Weighted-object ensemble clustering: Methods and analysis*, Knowledge Inf. Syst. **51** (2017), no. 2, 661–689.

[42] N.C. Sandes and A.L.V. Coelho, *Clustering ensembles: A hedonic game theoretical approach*, Pattern Recogn. **81** (2018), 95–111.

[43] C.P. Santos, D.M. Carvalho, and M.C.V. Nascimento, *A consensus graph clustering algorithm for directed networks*, Expert Syst. Appl. **54** (2016), 121–135.

[44] A.J.C. Sharkey, *Combining artificial neural nets: Ensemble and modular multi-net systems*, Springer Science & Business Media, 2012.

[45] J. Shi and J. Malik, *Normalized cuts and image segmentation*, IEEE Trans. Pattern Anal. Machine Intel. **22** (2000), no. 8, 888–905.

[46] R. Sibson, *Slink: An optimally efficient algorithm for the single-link cluster method*, Comput. J. **16** (1973), no. 1, 30–34.

[47] A. Strehl and J. Ghosh, *Cluster ensembles—a knowledge reuse framework for combining multiple partitions*, J. Machine Learn. Res. **3** (2002), no. Dec, 583–617.

[48] A. Topchy, A.K. Jain, and W. Punch, *Combining multiple weak clusterings*, Third IEEE Int. Conf. Data Min., IEEE, 2003, pp. 331–338.

[49] ______, *A mixture model for clustering ensembles*, Proc. SIAM Int. Conf. Data Min., SIAM, 2004, pp. 379–390.

[50] ______, *Clustering ensembles: Models of consensus and weak partitions*, IEEE Trans. Pattern Anal. Machine Intel. **27** (2005), no. 12, 1866–1881.

[51] S. Vega-Pons and J. Ruiz-Shulcloper, *A survey of clustering ensemble algorithms*, Int. J. Pattern Recogn. Artific. Intel. **25** (2011), no. 3, 337–372.

[52] X. Wang, D. Han, and C. Han, *Rough set based cluster ensemble selection*, Proc. 16th int. Conf. Inf. Fusion, IEEE, 2013, pp. 438–444.

[53] J. Wu, H. Liu, H. Xiong, J. Cao, and J. Chen, *K-means-based consensus clustering: A unified view*, IEEE Trans. Knowledge Data Engin. **27** (2014), no. 1, 155–169.

[54] X. Wu, T. Ma, J. Cao, Y. Tian, and A. Alabdulkarim, *A comparative study of clustering ensemble algorithms*, Comput. Electric. Engin. **68** (2018), 603–615.

[55] F. Yang, X. Li, Q. Li, and T. Li, *Exploring the diversity in cluster ensemble generation: Random sampling and random projection*, Expert Syst. Appl. **41** (2014), no. 10, 4844–4866.

[56] Z. Yu, L. Li, J. Liu, J. Zhang, and G. Han, *Adaptive noise immune cluster ensemble using affinity propagation*, IEEE Trans. Knowledge Data Engin. **27** (2015), no. 12, 3176–3189.

[57] Z. Yu and H.-S. Wong, *Class discovery from gene expression data based on perturbation and cluster ensemble*, IEEE Trans. Nanobiosci. **8** (2009), no. 2, 147–160.

[58] X. Zhao, F. Cao, and J. Liang, *A sequential ensemble clusterings generation algorithm for mixed data*, Appl. Math. Comput. **335** (2018), 264–277.

[59] L. Zheng, T. Li, and C. Ding, *A framework for hierarchical ensemble clustering*, ACM Trans. Knowledge Disc. From Data (TKDD) **9** (2014), no. 2, 1–23.

[60] S. Zhong and J. Ghosh, *A comparative study of generative models for document clustering*, Proc. Workshop Cluster. High Dimens. Data Appl. SIAM Data Min. Conf., Citeseer, 2003.