machine learning & knowledge extraction

MDPI

Review

Categorical Data Clustering: A Bibliometric Analysis and Taxonomy

Maya Cendana and Ren-Jieh Kuo *

Department of Industrial Management, National Taiwan University of Science and Technology, No. 43, Section 4, Kee-Lung Road, Taipei 106, Taiwan; d10901809@mail.ntust.edu.tw
* Correspondence: rjkuo@mail.ntust.edu.tw; Tel.: +886-2-27376328; Fax: +886-2-27376344

Abstract: Numerous real-world applications apply categorical data clustering to find hidden patterns in the data. The *K*-modes-based algorithm is a popular algorithm for solving common issues in categorical data, from outlier and noise sensitivity to local optima, utilizing metaheuristic methods. Many studies have focused on increasing clustering performance, with new methods now outperforming the traditional *K*-modes algorithm. It is important to investigate this evolution to help scholars understand how the existing algorithms overcome the common issues of categorical data. Using a research-area-based bibliometric analysis, this study retrieved articles from the Web of Science (WoS) Core Collection published between 2014 and 2023. This study presents a deep analysis of 64 articles to develop a new taxonomy of categorical data clustering algorithms. This study also discusses the potential challenges and opportunities in possible alternative solutions to categorical data clustering.

Keywords: categorical data clustering; *K*-modes algorithm; bibliometric analysis; taxonomy of clustering algorithm

check for updates

Citation: Cendana, M.; Kuo, R.-J. Categorical Data Clustering: A Bibliometric Analysis and Taxonomy. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 1009–1054. https://doi.org/10.3390/make6020047

Academic Editors: Liang Zhao, Liang Zou and Boxiang Dong

Received: 28 February 2024
Revised: 6 April 2024
Accepted: 27 April 2024
Published: 7 May 2024

1. Introduction

Currently, the internet and artificial intelligence are undergoing significant development. Consequently, they generate vast quantities of transaction data, including structured data such as personal biodata, surveys, stock market data, medical records, marketing data, and e-commerce transactions, as well as data generated from applications used in various fields such as science, engineering, or unstructured data gathered from the internet, such as data from the Google search engine or information extracted from social media platforms. Therefore, mining these data to derive insightful information has become more important. This process is called data mining or knowledge discovery in databases (KDD).

Numerous methods exist in data mining, depending on how they process the data. For example, supervised learning involves processing data based on their labeled attributes. This method utilizes historical data to train the model and subsequently generates results based on the patterns learned during training. Classification, prediction, and regression are tasks performed under supervised learning. In contrast, unsupervised learning involves processing data without explicit supervision or labeled target variables. One common method is clustering, which identifies hidden patterns based on similarities within the data. The more similar the data points are, the more likely it is they will be grouped into the same cluster.

Clustering finds applications in various real-world scenarios, including market segmentation [1,2], healthcare [3,4], image processing [5–7], bioinformatics [8,9], social sciences [10], and text mining [11].

Since similarity is an important factor in enhancing cluster quality, it is essential to comprehend how to measure the similarity between data objects in the dataset. Each data object possesses attributes, each distinguished by its data type. For instance, the Iris dataset [12] comprises 150 data objects and four attributes: sepal length, sepal width, petal

length, and petal width. These attributes are categorized as numerical data, while another category is categorical data. Numerical data types are further divided into interval and ratio, while categorical types are categorized into nominal and ordinal [13].

Different data types require different distance metrics to calculate the similarity between their data objects, and different clustering algorithms are employed to process them. Two widely-used clustering algorithms include the *K*-means algorithm [14], used for clustering numerical data, and the *K*-modes algorithm [15], utilized for clustering categorical data.

To distinguish the *K*-modes algorithm from the *K*-means algorithm, there exist at least three characteristics concerning similarity metrics and how they represent the cluster: (1) The *K*-means algorithm employs means to represent the clusters, whereas the K-modes algorithm uses modes. (2) The *K*-means algorithm utilizes Euclidean distance, whereas the *K*-modes algorithm employs a dissimilarity metric. (3) The *K*-modes algorithm applies a frequency-based method to update the mode. Indeed, these algorithms exhibit various variations, and numerous studies classify them into several taxonomies of clustering methods that complement each other.

A respected taxonomy of clustering methods was proposed by [16], which categorizes clustering into hierarchical and partitional based on how clusters are produced. Additionally, clustering can be differentiated based on memberships, such as hard and fuzzy clustering. Taxonomies may vary and overlap. For instance, ref. [17] categorizes clustering into hard, fuzzy, and rough set clustering. Another taxonomy, proposed by [18], classifies distance or similarity metrics for categorical data, distinguishing similarity into context-sensitive and context-free, with the context-free category comprising probabilistic, information-theoretic, and frequency-based approaches.

Furthermore, various other taxonomies and surveys exist, covering topics that focus on determining the cluster number [19], machine-learning-based clustering [20], big data clustering [21,22], density peak clustering [23], subspace clustering for high-dimensional data [24], automatic clustering [25], and how nature-inspired metaheuristic techniques are implemented into automatic clustering [26].

Previous review studies employ various methodologies such as systematic literature reviews (SLR), bibliometric analysis, or meta-analysis depending on the scope of the studies, number of studies, and objectives. Therefore, drawing inspiration from previous review studies, particularly those focusing on categorical data, as shown in Table 1, this study aims to develop a taxonomy to refine the previous classification of categorical data clustering. This objective will be pursued by performing bibliometric analysis, presenting quantitative and qualitative synthesis, and analyzing eligible articles based on content screening.

Table 1. Previous review studies in the clustering domain.

Author (Year)	Summary	Database	Data of Collection Period	#Articles/Algorithms	Methodology
* Alamuri et al. (2014)	A taxonomy of categorical data similarity measures and the categorical clustering algorithms [18]	Official website	n/a	n/a	Survey
Parsons et al. (2004)	A survey of the various subspace clustering algorithms [24]	Official website	n/a	11 algorithms	Survey
Hancer & Karaboga (2017)	A comprehensive review of the determination of cluster numbers based on traditional, merge-split, and evolutionary computation (EC)-based approaches [19]	Official website	n/a	**Single-objective:** 43 algorithms; **multi-objective:** 15 algorithms	Survey

Table 1. *Cont.*

Author (Year)	Summary	Database	Data of Collection Period	#Articles/Algorithms	Methodology
Alloghani et al. (2019)	A systematic review of supervised and unsupervised machine learning techniques [20]	EBSCO, ProQuest Central Databases	2015–2018	84 articles	SLR and Meta-analysis using PRISMA
Naouali et al. (2019)	A survey of categorical clustering. The classification of clustering consists of hard, fuzzy, and rough sets and three evaluation metrics (accuracy, precision, and recall) [17]	Official website	1998–2017	32 algorithms	Survey
Ezugwu (2020)	A taxonomical overview and bibliometric analysis of clustering algorithms and automatic clustering algorithms, as well as the systematic review of all the nature-inspired metaheuristic algorithms for both non-automatic and automatic clustering [25]	WoS database	1989–2019	4875 articles for bibliometric analysis and 86 articles for SLR (45 non-automatic and 40 automatic clustering)	SLR, Bibliometric analysis via VOSviewer
Ezugwu (2020)	A systematic review of nature-inspired metaheuristic algorithms for automatic clustering [26]	Scopus database	n/a–2020	1649 articles for bibliometric analysis; 37 automatic clustering algorithms; and an experimental study of 5 metaheuristic algorithms using 41 datasets	SLR, Bibliometric analysis via VOSviewer
Awad & Hamad (2022)	A review of clustering techniques to handle big data issues [21]	Publisher website	2015–2022	>200 articles	SLR
Ikotun et al. (2023)	A comprehensive overview and taxonomy of the *K*-means clustering algorithm and its variants [22]	Publisher website	1984–2021	83 articles	SLR
Wang et al. (2024)	A review of all the density peak clustering (DPC)-related works [23]	Google Scholar and WoS database	2014–2023	>110 articles	SLR
This study	An up-to-date taxonomical overview and bibliometric analysis for categorical data clustering	WoS database	2014–2023	64 articles	Combine PRISMA and bibliometric analysis procedure via VOSviewer

* Related to categorical data.

Figure 1 illustrates the research design procedure, which integrates the PRISMA guide [27] with the standard general science mapping workflow for conducting bibliometric analysis [28–31]. The bibliometric data are sourced from the Web of Science (WoS) Core Collection [32], and the analysis methodology relies on science mapping techniques, utilizing network analysis to visualize citation networks, including co-word and citation data [30,33,34]. Furthermore, the visualization of each network is represented in a two-dimensional map using the free bibliometric software VOSviewer [35,36], facilitating a better interpretation of the output.

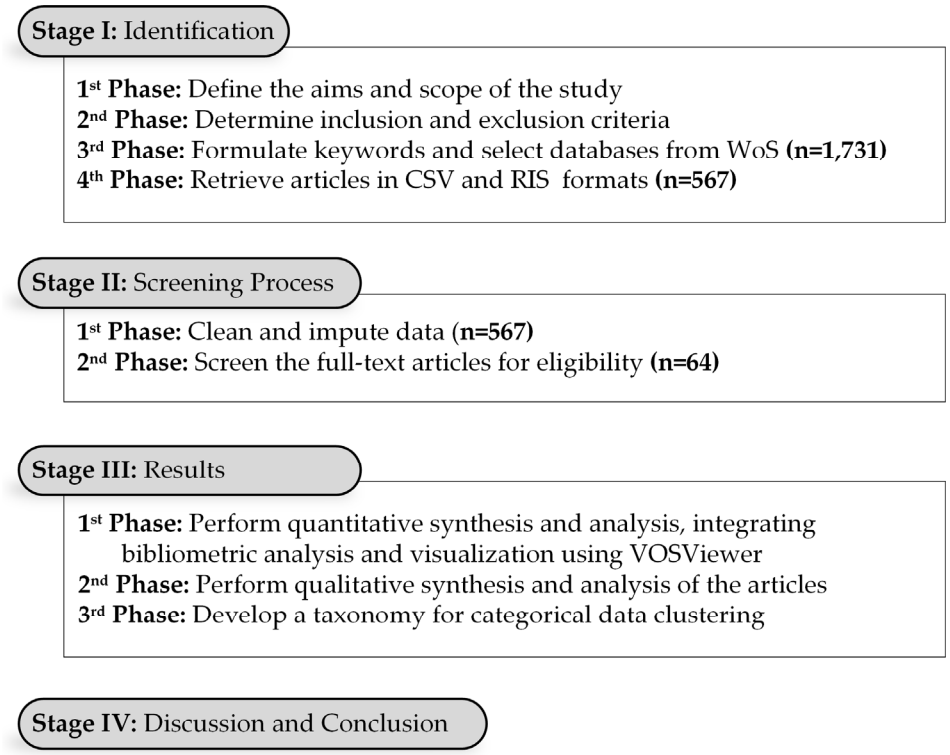

Figure 1. The research design procedure combines the PRISMA guides with the bibliometric analysis procedure.

The aims and scope of the studies are formulated into the following research questions:

1. What are the existing categorical data clustering algorithms capable of increasing clustering performances?
2. What are the research trends based on the co-word and citation network of these algorithms?
3. What is the updated taxonomy for categorical data clustering?
4. What potential challenges and future research directions exist as alternative solutions to the existing methods?

Therefore, this study contributes in the following ways: (1) by presenting a research-area-based bibliometric analysis based on over 50 articles published between 2014 and 2023; (2) by developing a new taxonomy of categorical data clustering algorithms and their variations; and (3) by discussing potential challenges and future research directions as alternative solutions to existing methods.

The remainder of this paper is structured as follows: Section 2 outlines the methods used in this study. Section 3 presents the analysis and develops a taxonomy of categorical data clustering. Section 4 discusses potential challenges and future research directions as alternative solutions to existing methods. Lastly, Section 5 provides the conclusions.

2. Methods

This study conducted a bibliometric analysis and developed a taxonomy. Additionally, it provides an overview of articles on categorical data clustering. The integration of quantitative and qualitative methods follows established guidelines for systematic reviews and meta-analysis, such as the PRISMA guide [27], alongside the standard general science mapping workflow for bibliometric analysis [28–31]. Figure 1 illustrates the research design procedure, which comprises four stages, each divided into several phases. The first stage, identification, involves defining the aims and scope of this study in its initial phase. Subsequently, the second, third, and fourth phases relate to data collection, as summarized in Table 1.

The second stage remains closely linked to the first, involving data cleaning and imputation to ensure no duplication. However, misspellings identified during data retrieval from the WoS, such as "k-modes," need correction to "k-modes." The next phase involves content analysis to determine the eligibility of articles based on the aims, scope, and criteria of this study.

In the third stage, the results are categorized into three phases: (1) quantitative synthesis and analysis, (2) qualitative synthesis and analysis, and (3) taxonomy development. In the first phase, bibliometric analysis is employed to analyze the performance of articles, focusing on publication years, titles, publishers, and authors. Moreover, visualization using co-word and citation networks is demonstrated for science mapping. Following this, the qualitative synthesis and analysis phase is conducted, resulting in the development of the taxonomy.

In the final stage, the discussion and conclusion are presented. This section addresses potential challenges and outlines future research directions, contributing alternative solutions to the existing methods.

In line with the previous study detailed in Table 1 and the research design procedure outlined in Figure 1, articles are retrieved in the CSV and RIS formats, and keywords and databases are specified. Given the emphasis on categorical data clustering, the chosen keywords are "clustering" and "categorical data." Although the terms "categorical data clustering" and "clustering categorical data" are often used interchangeably, they may convey slightly different connotations depending on the context. Generally, both phrases refer to the process of grouping or categorizing data points with categorical attributes.

The database used in this study is the WoS Core Collection. Apart from being well known for hosting high-quality journals, previous review papers in the clustering domain, as depicted in Table 1, have been analyzed regarding the databases employed for article retrieval. For instance, Ezugwu conducted two systematic reviews. In the first review, Ezugwu [25] utilized the WoS database to examine nature-inspired metaheuristics algorithms for both non-automatic and automatic clustering, identifying 40 automatic clustering algorithms. In the second review, Ezugwu [26] utilized the Scopus database and identified 37 automatic clustering algorithms. Notably, using only the WoS database is considered sufficient for retrieving clustering-related articles. Additionally, many review papers rely on the official website [17–19,24] or the publisher's website [20–22] as their primary source. In another study, Wang et al. [23] retrieved articles indexed by Google Scholar and the WoS database. Hence, after considering and comparing numerous database sources, this study completely restricts the database to the WoS Core Collection. A total of 1731 articles were identified. After applying the inclusion criteria, only 567 articles were selected for further analysis. These articles were directly retrieved from the WoS in CSV and RIS format.

Furthermore, the first criterion specifies publication years with index dates between 1 January 2014 and 5 December 2023. This time frame is selected to align with a previous survey on categorical data clustering conducted in [17]. That study represents the first survey on categorical data, analyzing 32 algorithms over 30 years, from 1998, when the *K*-modes algorithm was first introduced, to 2017. From 32 algorithms, 27 are related, and only 6 meet the inclusion criteria in this study [37–42].

The second criterion is the document types. This study excludes proceeding papers, book chapters, editorial materials, and other document types retaining only articles, as the number of other document types is insignificant. All documents are in English and belong to areas of computer science, mathematics, and engineering. Further details are provided in the flow diagram according to PRISMA 2020, shown in Figure 2. Additionally, Table 2 presents a summary of the keywords used and the databases selected, where *n* denotes the number of articles.

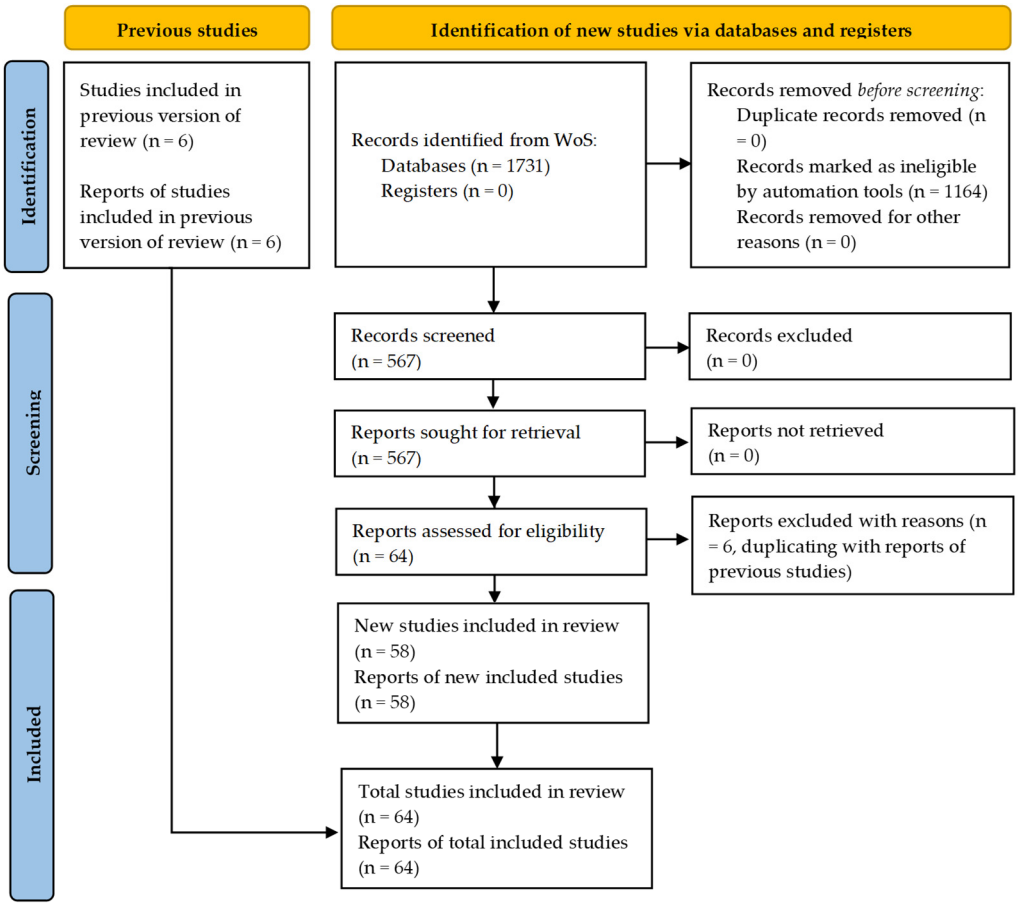

Figure 2. PRISMA flow diagram.

Table 2. Keywords and database selection.

Filters	n
Keywords: ((ALL = (clustering)) AND ALL = (categorical data))	1731
Publication Years: 2014–2023 (index date: 1 January 2014 to 5 December 2023	1113
Document Types: Article	1083
Languages: English	1067
Research Areas: Computer science, mathematics, and engineering	567
Content Screening	64

The inclusion criteria for articles involve the topic of categorical data clustering, focusing specifically on partition-based clustering and its variations. However, articles on statistic-based or model-based clustering, such as Latent Data Analysis [43–45] and EM algorithms [46], are excluded. Additionally, the exclusion criteria encompass articles related to multiview, co-clustering, consensus, deep learning clustering, and methods related to data stream clustering. This decision is based on the fact that many of these methods are employed in semi-supervised learning, which differs from the unsupervised learning approach adopted in this study.

Furthermore, this study excludes algorithms that process numerical, mixed (numerical and categorical), text data, and sequential categorical data. Despite the variety of data types, processing categorical data in clustering remains challenging compared to numerical data. This is primarily due to differences in calculating the distance between data points; therefore, the scope of this study is limited to addressing the specific research questions.

3. Results

This section consists of three phases: (1) quantitative synthesis and analysis, (2) qualitative synthesis and analysis, and (3) taxonomy development.

3.1. Quantitative Synthesis and Analysis

Quantitative synthesis and analysis involve employing bibliometric analysis to explore trends, identify patterns, and analyze the performance of articles. Initially, the contributions of research constituents related to categorical data clustering are assessed by presenting the performance analysis in a descriptive format. Three types of performance analysis are conducted: (1) citation-related metrics, (2) publication-related metrics, and (3) citation-and-publication-related metrics. This study solely utilizes publication-related metrics. Subsequently, science mapping is performed to investigate the relationships between research constituents. Co-word and citation analyses are employed to determine the connections between topics/keywords and cited publications.

3.1.1. Performance Analysis

In this subsection, the performance summary includes (1) publication years, (2) publication titles, (3) publishers, and (4) authors.

- Publication Years

Figure 3 shows the publication years, where the Y-axes represent the number of publications (left) and citations (right). The publication trend shows an increase, with 2019 emerging as the most productive year. In that year, 11 articles related to clustering were published, covering hierarchical-based [47,48], rough-set-based [49,50], weight-based [51], graph-based [52], a variant of fuzzy clustering [53–55], integer linear programming [56], and clustering validity [57].

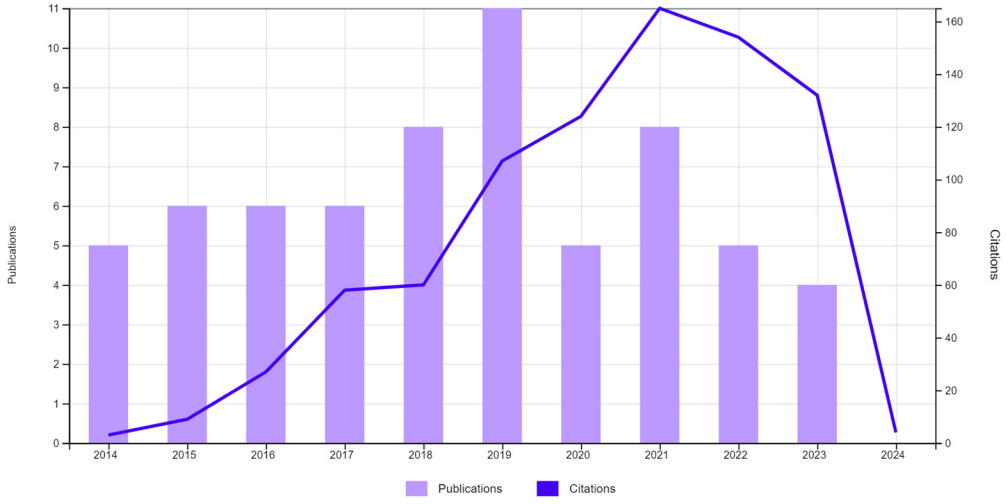

Figure 3. Publication years.

Moreover, for a deeper comprehension of the citations and publications spanning the period from 2014 to 2023, Table 3 presents the top ten cited articles. TC denotes the total citations, while AC represents the average citations per year.

- Publication Titles and Publishers

Table 4 shows the publication titles, with TP representing the total publications. All publication titles (journals) are categorized under "computer science, artificial intelligence" or "computer science, information systems." In total, there are 41 journals, with Neurocomputing ranking highest on the list. Furthermore, Table 5 presents the publishers, with these five publishers covering over 80% of published articles.

Table 3. Times cited.

Articles	AC	TC
A new distance metric for unsupervised learning of categorical data [58]	8.22	74
Initialization of *K*-modes clustering using outlier detection techniques [37]	7.33	66
Space structure and clustering of categorical data [59]	6	54
Hierarchical clustering algorithm for categorical data using a probabilistic rough set model [38]	3.64	40
Non-dominated sorting genetic algorithm using fuzzy membership chromosome for categorical data clustering [60]	3.5	35
Soft subspace clustering of categorical data with probabilistic distance [61]	3.67	33
Rough set approach for clustering categorical data using information-theoretic dependency measure [62]	3	30
Many-objective fuzzy centroids clustering algorithm for categorical data [63]	3.71	26
Comparison of similarity measures for categorical data in hierarchical clustering [48]	4.17	25
A fast and effective partitional clustering algorithm for large categorical datasets using a *K*-means based approach [64]	3.43	24

Table 4. Publication titles.

Publication Titles	TP
Neurocomputing	7 (10.938%)
IEEE Transactions on Neural Networks and Learning Systems	5 (7.813%)
Applied Soft Computing	3 (4.688%)
Pattern Recognition	3 (4.688%)
IEEE Access	3 (4.688%)
Applied Intelligence	2 (3.125%)
Engineering Applications of Artificial Intelligence	2 (3.125%)
Expert Systems with Applications	2 (3.125%)
Information Sciences	2 (3.125%)
International Journal of Machine Learning and Cybernetics	2 (3.125%)
Knowledge-based Systems	2 (3.125%)
Mathematics	2 (3.125%)
Others (29 publication titles)	29 (45.31%)

Table 5. Publishers.

Publishers	TP
Elsevier	27 (42.188%)
IEEE	13 (20.313%)
Springer Nature	11 (17.188%)
MDPI	3 (4.688%)
IOS Press	2 (3.125%)
Others (8 publishers)	8 (12.498%)

- Authors

Each author contributes a specific area of categorical data clustering. However, the total publication (TP) presented in Table 6 shows the total number of articles authored by both authors and co-authors. Ten authors contribute to more than 50% of the articles.

Furthermore, the most productive authors in categorical data clustering are J. Y. Liang and R. J. Kuo. These authors have collaborated with co-authors who rank highly and have contributed three to four articles. For example, J. Y. Liang, as a co-author, collaborated with L. Bai on a study optimizing the objective function of partition clustering [39,65,66]. Additionally, J. Y. Liang collaborated with F. Y. Cao and J. Z. X. Huang on proposing clustering algorithms tailored for various data types, such as set-valued features [40,67] and matrix-object data [68]. Collaborations with authors such as W. Wei [47] and Y. H. Qian [59] further exemplify J. Y. Liang's significant contributions to the field.

Table 6. Authors.

Authors	TP
J. Y. Liang	8 (12.500%)
R. J. Kuo	6 (9.375%)
S. B. Salem	4 (6.250%)
Z. Chtourou	4 (6.250%)
S. Naouali	4 (6.250%)
Y.M. Cheung	4 (6.250%)
T. P. Q. Nguyen	4 (6.250%)
L. Bai	3 (4.688%)
F. Y. Cao	3 (4.688%)
J. Z. X. Huang	3 (4.688%)
L. F. Chen	3 (4.688%)
Y. Q. Zhang	3 (4.688%)
Others (153 authors)	18 (29.685%)

Similar to J. Y. Liang, R. J. Kuo has conducted studies on metaheuristic-based clustering in collaboration with T. P. Q. Nguyen [2,53–55,60,69]. Other authors, such as S. Salem, S. Naouali, and Z. Chtourou, have also worked on rough-set clustering [64,70–72]. Furthermore, Y. M. Cheung, as the second author, has proposed numerous methods related to distance metrics with Y. Q. Zhang [73–75] and H. Jia [58]. Moreover, F. L. Chen proposed variant methods for optimizing the objective function in subspace clustering algorithms [61,76,77].

Considering the most productive authors alongside the times the articles were cited in Table 4, several findings are revealed: (1) four of the ten articles are attributed to the top ten authors [59–61,64]. However, it is notable that the first author of the most cited article is not the most productive, even though their co-author is included among the ten most productive authors. (2) Despite the productivity of certain authors, articles focusing on topics like set-valued features and matrix-object data, as well as the subject of cluster validity, do not appear to receive significant citation counts. (3) Many articles related to metaheuristic published after 2018 are also not among the top cited articles. Nevertheless, these findings warrant further investigation, primarily due to the scope and limitations of this study, including the range of publication years and the impact of topics on citation counts.

3.1.2. Science Mapping

Science mapping constitutes one of the principal methodologies in bibliometric analysis. It involves a range of techniques, each distinguished by its usage and data utilization. These techniques include citation analysis, co-citation analysis, bibliographic coupling, co-word analysis, and co-authorship analysis [29]. For the purposes of this study, we focused solely on two specific techniques: co-word analysis and citation analysis.

Co-word analysis involves examining the co-occurrence of word pairs or the frequency with which two or more words appear together in a given corpus. In this study, the words were extracted from "author keywords". This method operates on the assumption that keywords frequently appearing together are thematically related, thereby aiding in the formation of thematic clusters that define specific topics.

In contrast, citation analysis focuses on the relationships among publications rather than their content. Additionally, for further examination, co-citations can be employed to relate publications frequently cited together. In a co-citation network, the connection between two publications is determined by their co-occurrence in the reference lists of other publications. Although co-citation analysis can identify highly influential publications, this study primarily aims to explore the relationships among publications over a specific ten-year period. Consequently, the use of co-citations in this study might yield overly generalized results.

Moreover, visualization techniques assist as valuable tools for representing the science map. Each science map employs distinct analysis techniques and algorithms. Cobo [30] conducted a comparative study of nine science mapping applications, clarifying their advantages and drawbacks.

This study utilized VOSviewer, employing network analysis as its method. As illustrated in Figure 4, each label (keyword) is interconnected, with the size of the labels corresponding to their frequency. Bigger labels indicate a higher frequency of appearance. Furthermore, thematic clusters are distinguished through the use of different colors in the visualization. The color sequence is as follows: red, green, blue, and yellow.

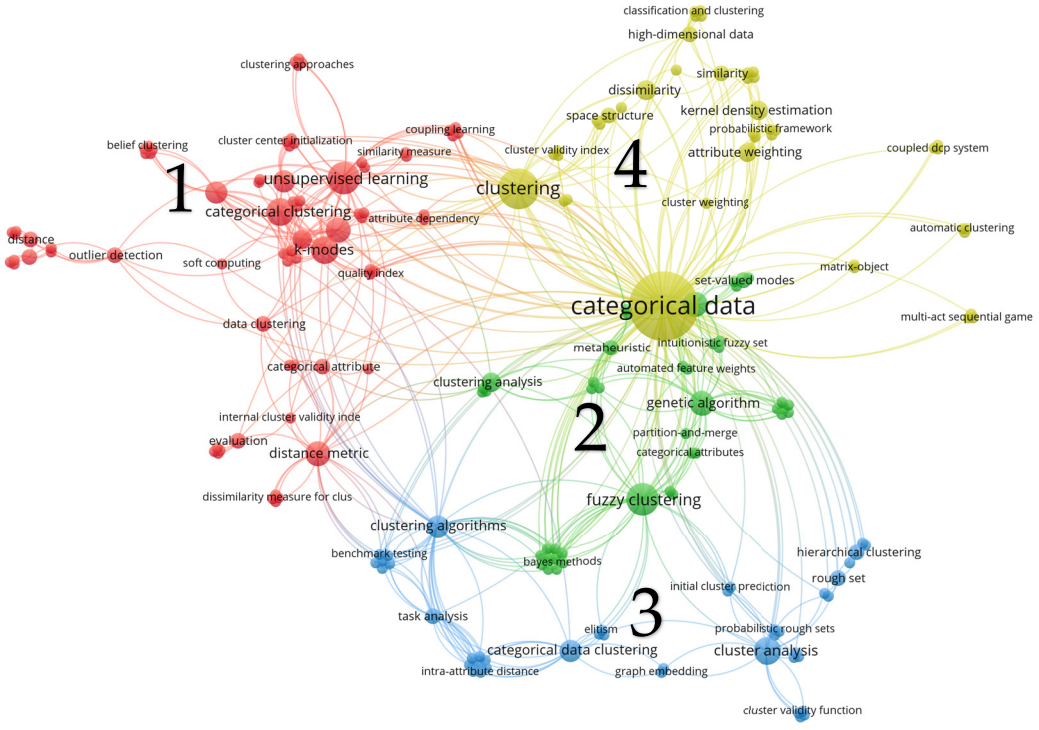

Figure 4. Co-word network.

The co-word network depicted in Figure 4 reveals four thematic clusters comprising a total of 182 keywords. Notably, Cluster 1 exhibits a significantly larger size compared to the other clusters. While certain keywords such as "distance metric," "internal cluster validity index," "evaluation," and "dissimilarity measure for clustering" possess a general scope, a few keywords stand out for their unique association with concepts such as "outlier detection," "k-modes," "k-modes clustering," "condorcet clustering," and "rough set theory."

Within Cluster 2, numerous keywords relating to variations of fuzzy concepts are apparent, including "fuzzy centroid," "fuzzy clustering," "fuzzy k-modes," "fuzzy k-modes algorithm," "fuzzy sv-k-modes," "intuitionistic fuzzy set," "rough fuzzy clustering," and "wfk-modes." Furthermore, alongside fuzzy clustering, keywords related to metaheuristics such as "genetic algorithm," "particle swarm optimization," "sine cosine algorithm," and

"simulated annealing" are prevalent. Additionally, another prominent topic within Cluster 2 is "multi-objective optimization."

Cluster 3 presents a distinct set of algorithms compared to the previous clusters. Specific keywords within this cluster include "hierarchical clustering," "graph embedding," "divisive clustering," "granular computing," "locality-sensitive hashing," "distribution approximation," and "holo-entropy." Notably, the keyword "hierarchical clustering" strongly connects to "rough set."

The final cluster also features specific keywords, mainly related to "high-dimensional data," "attribute weighting," "cluster weighting," "dissimilarity," "similarity," "distance measure," "coupled dcp system," and "kernel density estimation," in addition to clustering methods such as the "k-mw-modes algorithm" and "automatic clustering." A summary of the thematic cluster is provided in Table 7.

Table 7. Co-word analysis.

Cluster	#Keywords	Summary
1	73	A strong connection exists between the *K*-modes algorithm and rough set theory. Furthermore, rough sets are linked to outlier detection, which, in turn, is associated with the initial cluster centers. This linkage suggests that rough sets are utilized to address outliers in the *K*-modes algorithm arising from the random initialization of cluster centroids.
2	43	This cluster covers the fuzzy clustering algorithm, including variations such as the fuzzy *K*-modes (FKM) algorithm and rough fuzzy clustering. Additionally, the cluster highlights a growing trend in optimizing fuzzy clustering using metaheuristic-based algorithms. Consequently, future studies should delve deeper into investigating the optimization of fuzzy clustering, leveraging not only genetic algorithms and particle swarm optimization but also other metaheuristics to enhance algorithm performance.
3	42	This cluster covers hierarchical clustering and its relationship with rough set theory. Additionally, it includes keywords related to cluster analysis, such as graph embedding and cluster validity functions.
4	42	The keywords in this cluster are associated with dissimilarity methods and attribute weighting, such as kernel density estimation and probabilistic frameworks.

Since co-word analysis relies on authors' keywords, redundancy can occur. Therefore, this study experimented with visualization techniques employing multiple clusters to address this issue. The findings revealed that four thematic clusters effectively identified and represented the relationships between categorical data clustering topics. Additionally, to clarify the relationships among publications, this study constructed a citation network. Among the 64 articles analyzed, 56 were interconnected, while 8 exhibited no connections.

The citation analysis shown in Figure 5 illustrates the relationships among publications, with bigger labels (articles) indicating the most influential publications. An interesting aspect to explore is the relationship between the total number of citations (TC) and the number of citations between publications (links). For example, reference [58] by Jia et al. in 2016 is associated with 15 links and 74 TCs. In other words, out of the 64 articles analyzed, 15 are linked to the work of Jia et al. [58]. Further details are provided in Table 8.

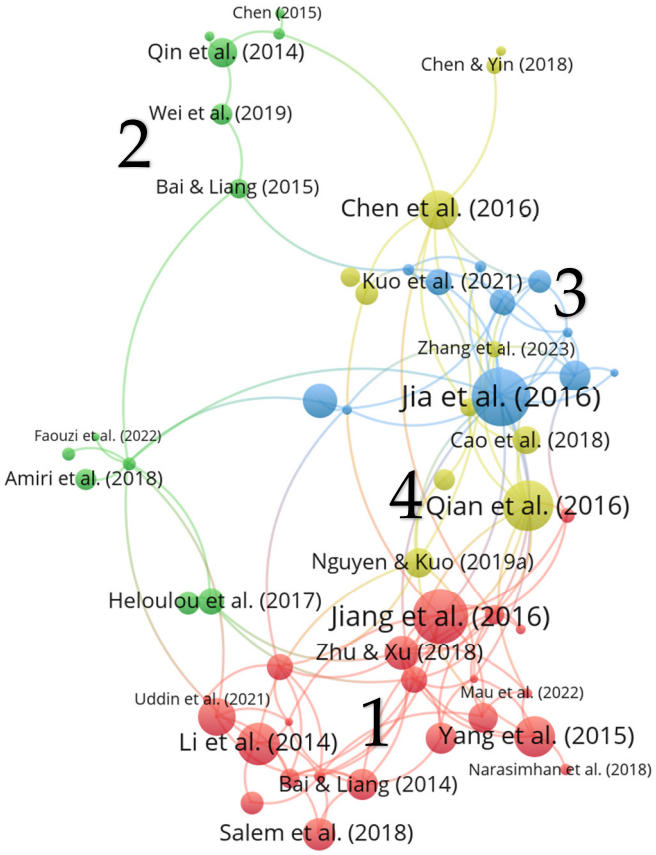

Figure 5. Citation network.

Table 8. Citation analysis.

Articles	Cluster	Links	TC	Articles	Cluster	Links	TC
Jia et al. (2016) [58]	3	15	74	Yuan et al. (2020) [78]	1	4	5
Oskouei et al. (2021) [79]	1	11	8	Saha & Das (2015) [80]	1	4	22
Zhu & Xu (2018) [63]	1	10	26	Heloulou et al. (2017) [81]	2	4	15
Salem et al. (2021) [71]	1	9	3	Xiao et al. (2019) [56]	4	4	12
Kuo & Nguyen (2019) [55]	1	9	15	Salem et al. (2018) [64]	1	3	24
Jiang et al. (2016) [37]	1	9	66	Chen et al. (2021) [76]	2	3	3
Dorman & Maitra (2022) [82]	2	9	4	Bai & Liang (2015) [65]	2	3	9
Qian et al. (2016) [59]	4	9	54	Qin et al. (2014) [41]	2	3	19
Chen et al. (2016) [61]	4	8	33	Rios et al. (2021) [83]	3	3	3
Yanto et al. (2016) [42]	1	7	15	Cao et al. (2018) [40]	4	3	16
Bai & Liang (2014) [39]	1	7	22	Uddin et al. (2021) [84]	1	2	1
Nguyen & Kuo (2019a) [54]	4	7	19	Peng & Liu (2019) [51]	1	2	2
Naouali et al. (2020) [72]	1	6	9	Suri et al. (2016) [85]	1	2	11
Yang et al. (2015) [60]	1	6	35	Wei et al. (2019) [47]	2	2	10
Li et al. (2014) [38]	1	6	40	Kuo et al. (2021) [69]	3	2	15
Kar et al. (2023) [86]	3	6	2	Ye et al. (2019) [52]	3	2	0
Bai & Liang (2022) [66]	3	6	3	Chen & Yin (2018) [87]	4	2	6
Jian et al. (2018) [88]	3	6	20	Cao et al. (2017b) [67]	4	2	10

Table 8. *Cont.*

Articles	Cluster	Links	TC	Articles	Cluster	Links	TC
Zhang et al. (2023) [89]	4	6	6	Narasimhan et al. (2018) [90]	1	1	3
Zheng et al. (2020) [91]	4	6	7	Faouzi et al. (2022) [92]	2	1	0
Jiang et al. (2023) [93]	1	5	1	Nguyen & Kuo (2019) [53]	2	1	11
Salem et al. (2021a) [70]	1	5	0	Amiri et al. (2018) [94]	2	1	10
Park & Choi (2015) [62]	1	5	30	Kim (2017) [95]	2	1	4
Zhang & Cheung (2022a) [73]	3	5	2	Sun et al. (2017) [96]	2	1	2
Zhang & Cheung (2022b) [74]	3	5	11	Chen (2015) [77]	2	1	2
Zhang et al. (2020) [75]	3	5	15	Sulc & Rezankova (2019) [48]	3	1	25
Mau et al. (2022) [97]	1	4	2	Gao & Wu (2019) [57]	4	1	2
Dinh & Huynh (2020) [98]	1	4	18				

3.2. Qualitative Synthesis and Analysis

Qualitative synthesis and analysis will be conducted following the quantitative synthesis and analysis. This phase comprehensively explains the 64 articles identified through the screening process. First, the articles will be categorized according to the classification proposed by [16,17] which distinguishes between hierarchical clustering and partitional clustering. Partitional clustering further encompasses hard, fuzzy, and rough-set-based clustering methods. Subsequently, the third and fourth sections will focus on algorithms that specifically modify the distance function and weighting method, while the fifth section will discuss algorithms related to validity functions. Additionally, this subsection will provide a summary of the datasets and performance evaluation criteria utilized by the various algorithms. Detailed explanations and patterns identified during the analysis of these algorithms will be presented in the following subsections.

3.2.1. Hierarchical Clustering

The hierarchical clustering algorithms are categorized into divisive and agglomerative hierarchical clustering. Among the identified articles, three algorithms are based on divisive hierarchical clustering, while two focus on agglomerative hierarchical clustering, as shown in Table 9. Notably, many of these algorithms are based on information theory. Furthermore, there has been significant advancement in the performance of previous algorithms, shown by the improvement of the min-min-roughness (MMR) algorithm [99].

Table 9. Hierarchical clustering.

Authors (Year)	Algorithms	Methods	Comparisons
Li et al. (2014)	MDP, TMDP, MTMDP [38]	Divisive, based on probabilistic rough set theory approach	MMR
Qin et al. (2014)	MGR [41]	Divisive, based on an information-theoretic approach	MMR, *k*-ANMI [100], G-ANMI [101], COOLCAT [102]
Wei et al. (2019)	KOF, MNIG [47]	Divisive, based on an information-theoretic approach	MMR, MGR, MDA [103], TR [104]
Sun et al. (2017)	HPCCD [96]	Agglomerative, based on an information-theoretic approach	MGR [41], COOLCAT, LIMBO [105], *K*-modes,
Altameem et al. (2023)	P-ROCK [106]	Agglomerative, linked-based	ROCK [107]

(1) Divisive Hierarchical Clustering

Li et al. [38] introduced the maximum total mean distribution precision (MTMDP), aiming to improve the min-min-roughness (MMR) algorithm [99] based on probabilistic

rough set theory. The MTMDP algorithm involves three main improvements: (1) It utilizes distribution approximation precision instead of the accuracy of approximation employed in the MMR algorithm. (2) Candidate attributes are ranked by total mean distribution precision rather than by max mean distribution precision. (3) Leaf node splitting is performed based on the smallest cohesion degree rather than selecting the leaf node with more objects for further splitting clustering. As a result, the proposed algorithm demonstrates efficacy in handling uncertain and imbalanced datasets, enabling automatic cluster detection and enabling an analysis of high-dimensional datasets. A future study of the MTMDP algorithm may explore automatic subspace clustering for high-dimensional data or its implementation for mixed numeric and categorical datasets.

Similar to MTMDP, Qin et al. [41] introduced an algorithm inspired by MMR called mean gain ratio (MGR), which is based on information theory. Unlike the MMR algorithm, MGR avoids bias towards extreme selection, as extreme selection can potentially decrease accuracy. First, MGR selects a clustering attribute using the mean gain ratio and then identifies an equivalence class on the clustering attribute using cluster entropy. Notably, the MGR algorithm can operate without specifying the number of clusters. In each iteration, a cluster is discovered regardless of length, followed by a binary split on the remaining objects.

Consequently, this algorithm is well suited for large categorical datasets with imbalanced distributions. Experimental results demonstrate that MGR is efficient and scalable. In the future, enhancing accuracy can be pursued in two ways: integrating the MGR algorithm with the genetic clustering algorithm (G-ANMI) [101], and incorporating the reprocessing procedure from the COOLCAT algorithm [102].

Wei et al. [47] proposed another approach to improve the splitting of clusters in divisive hierarchical clustering. Initially, they conducted a comprehensive analysis of existing divisive hierarchical clustering algorithms, including MMR [99], MGR [41], MDA [103], and TR [104]. After that, they created a unified framework based on the strengths and weaknesses of these algorithms. Within this framework, the mean normalized information gain (MNIG) was introduced, specifically designed to address the limitations of MGR. Additionally, the *K*-modes object function (KOF) identifies suitable measures for attribute selection. Both KOF and MNIG contribute to determining the method for splitting clusters into subclusters and identifying which cluster should be split in each iteration. While KOF, MNIG, and other measures, such as a maximum number of objects (MO) and information entropies (IE), perform well in certain steps, identifying the optimal measure that universally fits all problems remains challenging.

(2) Agglomerative Hierarchical Clustering

On the other hand, Sun et al. [96] developed their algorithm based on agglomerative hierarchical clustering. Their proposed algorithm, named hierarchical projected clustering for categorical data (HPCCD), clusters high-dimensional data using the weighted holoentropy [108] instead of pairwise-similarity-based measures for merging two subclusters. HPCCD can distinguish relevant attributes within clusters and identify both the principal feature space and the core feature space, which is critical for clustering high-dimensional data. The experimental results indicate that HPCCD outperforms the MGR [41], *K*-modes [15], COOLCAT [102], and scalable information bottleneck (LIMBO) [105] in terms of efficiency, accuracy, and reproducibility.

In contrast to the aforementioned variations of hierarchical-based clustering, the algorithm proposed by Altameem et al. [106] stands out. Their approach aims to modify the ROCK algorithm [107] by allowing user-defined parameters as input, thus enhancing the flexibility of the algorithm. This modified version is named The Parameterized-ROCK (P-ROCK). The parameters involved include the threshold (θ) for neighborhood decision, f(θ), and h(θ). The P-ROCK algorithm was tested using two datasets from the UCI repository: the small soybean dataset and the congressional votes dataset. The results indicate that the P-ROCK algorithm shows improved accuracy and runtime compared to the original ROCK algorithm. Furthermore, P-ROCK outperforms other variations of ROCK, such as QROCK [109] and MROCK [110], in terms of computing time.

3.2.2. Partition Clustering

Partition clustering includes hard, fuzzy, and rough-set clustering methods [17]. Among the 64 articles analyzed, 11 developed algorithms based on hard clustering, 12 articles focused on fuzzy clustering, and 10 articles based on rough-set clustering. Each article is categorized based on its specific approach or characteristics, which helps identify its contributions.

(1) Hard Clustering

The summary of algorithms for hard clustering is presented in Table 10. Given the variation in terms and acronyms used across different articles or algorithms, this study standardized their names to enhance clarity. For instance, acronyms such as "KMD," "KM," "*K*-modes," and "Huang's *K*-modes" are all standardized as "*K*-modes" algorithms. However, it is worth noting that similar acronyms may refer to different algorithms; for instance, "WKM" and "Cao" may each refer to more than one algorithm. Additionally, there are cases where the same algorithm is referenced differently, such as the hamming distance (HD). In such cases, this study follows the conventions established in the original articles. For further details, refer to their corresponding references.

Table 10. Hard clustering.

Authors (Year)	Algorithms	Methods	Comparisons
Hariz & Elouedi (2014)	BCDP: IKBKM and DKBKM [111]	dynamic clustering based on the *K*-modes algorithm that uses the Transferable Belief Model (TBM) concepts [112]	BKM [113]
Cao et al. (2017)	*k*-mw-modes [68]	clustering categorical matrix-object data based on the *K*-modes algorithm	*K*-modes, W*k*-modes [114], Cao [115], FCCM [116]
Heloulou et al. (2017)	MOCSG [81]	the multi-objective clustering based-sequential game theoretic that extends the ClusSMOG algorithm [117]	*K*-modes, PAM [118], and single linkage algorithm [16]
Salem et al. (2018)	MF*k*-M [64]	frequency-based method to update the modes and the Manhattan distance metric to compute the distance	*K*-modes, *K*-means
Cao et al. (2018)	SV-*k*-modes [40]	heuristic method to update the centroids and the Jaccard coefficient to measure the distance between two set-valued objects	a multi-instance clustering algorithm (BAMIC) [119], *K*-modes, and Tr*K*-means [120]
Xiao et al. (2019)	IPO-ILP-VNS [56]	integer linear programming (ILP) approach under variable neighborhood search (VNS) framework	*K*-modes, Khan [121], *k*-MODET [37], Wu [122]
Dinh & Huynh (2020)	*k*-PbC [98]	the MFI-based approach integrated with partitional clustering with a kernel-based method and information-theoretic-based dissimilarity measure	*K*-means++ [123], *K*-means|| [124], Cao [125], Khan, *k*-MODET [37], *K*-modes, *K*-representatives [126], M-*K*-Centers (Mod-2) [127] and New (Mod-3) [128] and CD-Clustering [129]
Chen et al. (2021)	SKSCC [76]	subspace clustering algorithm based on kernel density estimation (KDE), self-expressiveness-based methods, and probability-based similarity measurement	*K*-modes, WKM [130], MWKM [131]
Bai & Liang (2022)	CDC_DR, CDC_DR + SE [66]	graph-based representation method	graph-embedding methods: Non-Embedding (NE), Spectral Embedding (SE) [132], Non-negative Matrix Factorization (NMF) [133], and Autoencoder (AE) [134] using joint and mean operation; categorical data encodings: *K*-modes, *K*-means with ordinal encoding, one-hot encoding [135], link-graph encoding [136], and coupled data embedding (CDE) [137]
Dorman & Maitra (2022)	OTQT [82]	based on hartigan algorithm for *K*-means algorithm	*K*-modes
Faouzi et al. (2022)	α-Condorcet [92]	based on Condorcet clustering [138]	*K*-modes

- Clustering various data types

Cao et al. conducted studies on different types of features. One study focused on a matrix-object with a one–many relationship, where one object has multiple feature vectors. Another study addressed set-valued features, where features can possess multiple values for an object—such as a person with multiple job titles and hobbies. To handle categorical matrix-object data, Cao et al. proposed the *K*-multi-weighted-modes (*k*-mw-modes) algorithm [68]. This algorithm introduces a new dissimilarity measure to compute the distance between two categorical matrix-objects and utilizes a heuristic approach to select the cluster center.

Similarly, their other proposed algorithm, the set-valued *K*-modes (SV-*k*-modes) algorithm [40], designed for clustering data with set-valued features, employs a heuristic approach to update cluster centers. The distance between two set-valued objects is measured using the Jaccard coefficient. Furthermore, this approach is tested for scalability and enhances the initialization mechanism for cluster centers. The results demonstrate a superior performance compared to benchmark algorithms, confirming that the SV-*k*-modes algorithm is scalable for large, high-dimensional datasets.

- Optimizing the number of clusters

One method for handling uncertainty is belief clustering for dynamic partition (BCDP), proposed by Hariz. This study extends the belief *K*-modes method (BKM) [113] to dynamic environments. Unlike BKM, which maintains a fixed number of clusters, objects, and features, BCDP considers the uncertainty of attribute values and the potential adjustment of cluster numbers using the concepts of cluster cohesion and separation concepts. This adjustment can involve either increasing (IK-BKM) [139] or decreasing (DK-BKM) [140] the number of clusters. As a result, the partitioning of clusters is updated without requiring a complete re-clustering process from scratch.

- Optimizing the cluster centers

In addition to determining the number of clusters beforehand, another obstacle faced by the *K*-modes algorithm is overcoming the initialization problem. Various algorithms have been developed based on dissimilarity measures, such as the optimal transfer quick transfer (OTQT) algorithm. The OTQT algorithm [82], developed by Dorman and Maitra, incorporates the Hartigan algorithm for the *K*-means algorithm [141]. Following the initialization step, the OTQT algorithm implements optimal and quick transfer stages to enhance the objective function rather than relying solely on distance metrics. One improvement in this method is ensuring that clusters are nonempty at the initialization step and in any iteration by initializing with *K* distinct modes. The OTQT algorithm demonstrates significantly improved accuracy and scalability in clustering complex data.

Dinh and Huynh [98] introduced a method for generating initial clusters based on frequent pattern mining, marking the first attempt to combine this approach with partitional clustering. The pattern-based clustering algorithm for categorical data (*k*-PbC) relies on the Fp-Max algorithm [142] for maximal frequent itemsets mining (MFIM). Additionally, *k*-PbC establishes cluster centers through a kernel density estimation method and computes distances using an information-theoretic-based dissimilarity measure (ITBD).

Chen et al. [76] also addressed the sensitivity of the *K*-modes algorithm in initializing clusters and modes by employing kernel clustering. They utilized the self-expressive kernel density estimation (SKDE) to develop a self-expressive kernel subspace clustering algorithm for categorical data (SKCC). SKCC incorporates feature weighting to discern the importance of attributes.

- Optimizing the objective function for large datasets

Fauzi et al. proposed the α-Condorcet [92] as an extension of Condorcet clustering [138]. Unlike the traditional approach of setting the number of clusters a priori using pairwise comparisons and a simple majority decision rule to maximize Condorcet's criterion, the α-Condorcet sets the number of clusters, α, beforehand. It introduces a new

Condorcet criterion function that incorporates similarity measures and proposes a heuristic algorithm. As a result, the algorithm efficiently processes large datasets and produces superior partitions compared to the *K*-modes algorithm for various values of α.

Clustering large-size datasets containing more than 100,000 data objects poses challenges in clustering categorical data. To address this issue, Xiao et al. [56] proposed a new algorithm that combines *K*-modes with integer linear programming (ILP). While ILP techniques are typically effective for small-size data, the proposed method leverages ILP and the framework of variable neighborhood search (VNS) to develop a heuristic approach. This approach minimizes the total inner-distance function of the *K*-modes algorithm, thereby reducing the computation cost of clustering large datasets.

- Optimizing the objective function based on a multi-objective approach

Another method related to multi-objective clustering based on sequential games is the MOCSG [81]. Inspired by their previous work, clustering based on sequential multi-objective games (CluSMOG) [117], MOCSG extends this approach to numerical data. As a multi-objective clustering algorithm, MOCSG integrates multiple objective functions to optimize R-square (RSQ), connectivity, and intra-cluster inertia objectives. Additionally, MOCSG can dynamically determine the number of clusters.

- Representing data based on the discretization method

Another method designed to handle large datasets is the Manhattan frequency *K*-means (MF*k*-M) algorithm [64], proposed by Ben Salem et al. MF*k*-M employs a *K*-means-based approach to process categorical data by converting it into numeric values using relative frequency. The use of relative frequency aims to improve the simple matching similarity measure [143]. Additionally, the algorithm utilizes Manhattan distance (L_1 norm) instead of Euclidean distance to address outliers and noisy data [144,145]. By adopting this approach, MF*k*-M results in lower computational costs than the *K*-modes algorithm, as computing means is less expensive than computing modes.

Similar to MF*k*-M, the algorithm proposed by Bai and Liang [76], categorical data clustering based on data representation with spectral embedding (CDC_DR+SE), also employs a conversion method to represent categorical data as a graph representation instead of using direct ordinal or one-hot encoding methods. The algorithm learns the representation of categorical values from their graph structure, easing the capturing of potential similarities between categorical values and their conversion into numerical data. Consequently, existing numerical clustering algorithms can effectively cluster categorical data.

(2) Fuzzy Clustering

A summary of the fuzzy clustering is presented in Table 11.

- Heuristic approach to cluster set-valued attributes

Fuzzy clustering offers fuzzy membership, allowing one object to belong to more than one cluster based on the percentage of membership. However, both hard and fuzzy clustering algorithms encounter similar challenges, and various techniques have been proposed to address their drawbacks. Cao et al. introduced the SV-*k*-modes algorithm for clustering categorical data with set-valued attributes [40] and extended it to fuzzy-based clustering, named fuzzy SV-*k*-modes [67].

- Multivariate membership approach

Furthermore, in relation to fuzzy membership, Maciel et al. introduced multiple fuzzy partitions for FKM to address the ambiguity in data that share properties across different clusters. Their proposed method, the multivariate fuzzy *K*-modes (MFKM) algorithm [148], acknowledges that attributes in distinct clusters may possess varying degrees of membership. This approach to membership assignment differs from FKM, which assigns uniform membership to all attributes across all clusters. Additionally, the study proposed an internal validation index termed the multivariate fuzzy silhouette index, capable of

assessing clustering validity by identifying a relevant subset of variables. Experimental results demonstrate that the MFKM algorithm yields superior solutions, particularly as the number of categories for each variable increase.

Table 11. Fuzzy clustering.

Authors (Year)	Algorithms	Methods	Comparisons
Yang et al. (2015)	NSGA-FMC [60]	fuzzy genetic algorithm and multi-objective optimization	GA-FKM [146], MOGA [147]
Cao et al. (2017)	Fuzzy SV-*k*-modes [67]	FKM for clustering the set-valued attributes	FKM
Maciel et al. (2017)	MFKM [148]	FKM with multivariate approach	FKM and LFkM [149]
Kuo et al. (2018)	PSOFKM, GAFKM, ABCFKM [2]	FKM with PSO, GA, and ABC algorithm	FKM
Narasimhan et al. (2018)	EGA-FMC [90]	GA-FKM with multi-objective rank-based selection	MOGA, GA-FKM, NSGA-FMC
Zhu & Xu (2018)	MaOFCentroids [63]	many-objective clustering with fuzzy centroid algorithm	FKM, Fuzzy Centroids [150], SBC [59], NSGA-FMC
Nguyen & Kuo (2019)	PM-FGCA [54]	MOGA with fuzzy membership chromosomes	*K*-Modes, FKM, GA-FKM, NSGA-FMC
Nguyen & Kuo (2019)	AFC-NSPSO [53]	automatic fuzzy clustering using non-dominated PSO	AT-DC [151], DHCC [152], PROCAD [153], MOCSG [81]
Kuo & Nguyen (2019)	IWFKM, GIWFKM [55]	intuitionistic fuzzy set and genetic algorithm	FKM, WFKM [80], GA-FKM, SBC, MaOFCentroids
Kuo et al. (2021)	PFKM, GA-PFKM, PSO-PFKM, SCA-PFKM [69]	possibilistic fuzzy *c*-means for the categorical data and metaheuristic methods (GA, PSO, and SCA)	FKM
Mau et al. (2022)	LSHF*K*-centers. [97]	locality-sensitive hashing (LSH)-based approach	FCM, FE*K*-means [154], SBC, *K*-medoids [155], *K*-modes, *K*-representative [126], *K*-centers [150],, F*K*-centers [156], FKM, SGA-Dist, SGA-Sep, SGA-SepDist [146], MOGA, NSGA-FMC, MaOFCentroids, LSH*K*-reps [157]
Jiang et al. (2023)	KIWFKM, KIWFKM-DCP [93]	intuitionistic fuzzy set and coupled DCP system	MEC [158], FSC [159], FKM, WFKM, IWFKM [55], GIWFKM [55]

- Metaheuristic approach

Another concern arises from the random initialization of centroids, leading to fast convergence to local optima. To address this, Kuo and Nguyen introduced metaheuristic-based fuzzy clustering to determine initial centroids, emphasizing global search. In their work [2], Kuo and Nguyen integrated the particle swarm optimization algorithm (PSO), genetic algorithm (GA), and artificial bee colony algorithm (ABC) with FKM. Among these methods, the GA-based FKM algorithm achieves the highest accuracy, with PSO demonstrating the most stability.

- Possibilistic-based approach with metaheuristic

Another study by Kuo et al. extended the possibilistic fuzzy *C*-means (PFCM) [160] to cluster categorical data, known as the possibilistic fuzzy *K*-modes (PFKM) algorithm. This

algorithm aims to overcome noise and outliers by employing frequency probability-based distance [58] as a dissimilarity measure and the possibility concept from the PFCM algorithm. After that, metaheuristic approaches are utilized to optimize the PFKM algorithm to achieve the optimal solution. Among the three methods considered—PSO, Sine-Cosine Algorithm (SCA), and GA—PFKM based on PSO and SCA demonstrates higher performance and requires less computational time compared to GA, which requires more complex updating rules.

- Intuitionistic fuzzy set theory-based approach

In [55], Kuo and Nguyen further integrated the frequency-probability-based distance metric with the intuitionistic fuzzy set (IFS), designed to handle uncertainty. Their study primarily extends previous methodologies employing IFS to cluster numerical datasets [161–164] to accommodate categorical data. Additionally, the study introduces attribute weighting, adopting the approach outlined by Saha and Das [80] within the framework of IFS, assigning weight factors to each categorical attribute.

However, the performance of the proposed method, the intuitionistic weighted fuzzy *K*-modes (IWFKM) algorithm by Kuo and Nguyen, is comparatively lower than benchmark algorithms, GA-FKM [146], SBC [59], and MaOfCentroids [63], due to the inability of IWFKM to prevent the local optima problem. Hence, to address this limitation, the authors propose a second algorithm, GIWFKM, which combines IWFKM with GA. The results demonstrate that GIWFKM outperforms all benchmark algorithms.

In 2023, Jiang et al. introduced an algorithm named the kernel-based intuitionistic weight fuzzy *K*-modes (KIWFKM), which integrates the IFS with kernel-trick and weighting mechanisms [93]. This algorithm aims to overcome noise and distinguish important attributes. Moreover, KIWFKM establishes the coupled DCP system, a chained tissue-like P system integrating DNA genetic rules. The P system, originally proposed by Paun [165], belongs to membrane computing, a nature-inspired computational model that can be optimized using a DNA genetic algorithm [166]. Consequently, KIWFKM is combined with the Coupled DCP system, as it provides a novel dynamic evolution model for existing P systems and can address non-combinatorial optimization problems. Experimental results demonstrate that the KIWFKM-DCP algorithms outperform other related algorithms across various datasets in terms of adjusted rand index (ARI), normalized mutual index (NMI), accuracy, and *F*-measure.

- Multi-objective approach

Furthermore, another algorithm based on fuzzy centroids, named MaOfCentroids [63], was proposed by Zhu and Xu. Their preliminary experiments suggest that fuzzy centroids are more effective and stable compared to other traditional fuzzy clustering. However, similar to other single-objective algorithms that suffer from finding the optimal partition, MaOfCentroids adopts a multi-objective clustering approach utilizing a reference point-based non-dominated sorting genetic algorithm to address this challenge. In this approach, fuzzy memberships serve as the chromosome representation. This study is significant as it was the first to employ more than three objective functions based on various cluster validity indexes (CVIs) to evaluate the specific structure or distribution of data.

Additionally, several other multi-objective clustering approaches have been integrated with fuzzy clustering, besides MaOfCentroids, include NSGA-FMC [60], EGA-FMC [90], AFC-NSPSO [53], and PM-FGCA [54]. In 2015, Yang et al. proposed the non-dominated sorting genetic algorithm-fuzzy membership chromosome (NSGA-FMC). NSGA-FMC aims to optimize clustering quality using fuzzy compactness and separation as objective functions. Unlike using attributes, NSGA-FMC initializes its chromosome with fuzzy memberships, thereby proposing a more efficient solution selection procedure that chooses a solution from the non-dominated Pareto front, leading to faster computation.

On the contrary, an enhanced genetic algorithm-based fuzzy *K*-modes clustering (EGA-FMC) proposed by Narasimhan [90] is derived from GA-FKM to enhance both the selection and elitism phases. Unlike the previous algorithm NSGA-FMC, EGA-FMC demonstrates

efficient clustering of larger datasets. Although the objective functions remain the same as NSGA-FMC, EGA-FMC employs multi-objective rank-based selection alongside enhanced elitism operations, ensuring the replacement of the worst child of the new population with the best parent before evolution.

Another way to approach multiple objectives is through automatic fuzzy clustering, using the non-dominated sorting particle swarm optimization (AFC-NSPSO) algorithm [53]. This algorithm aims for global compactness and fuzzy separation as objective functions. Moreover, the algorithm process is divided into two parts, incorporating control variables to automatically determine the cluster number and allocate objects to their respective clusters. Additionally, the proposed algorithm can identify the maximum number of clusters, which reduces computational time by minimizing iterations.

The main focus of multi-objective clustering algorithms is to enhance the performance of categorical data clustering according to the specific constraints of these algorithms. Another algorithm, known as the partition-and-merge-based fuzzy genetic clustering algorithm (PM-FGCA), is particularly dedicated to determining the optimal number of clusters within a predetermined number of clusters [54]. Initially, PM-FGCA employs a multi-objective fuzzy clustering approach similar to that of NSGA-FMC to generate an intermediate clustering solution based on the initial number of clusters. Subsequently, fuzzy centroids are utilized to improve the results. This process involves iteratively merging clusters until satisfactory solutions are obtained. Consequently, the computational time required by PM-FGCA tends to be longer compared to NSGA-FMC.

- Soft subspace clustering based on locality-sensitive hashing (LSH)

Mau et al. introduced the LSHF*k*-centers [97] algorithm, which incorporates locality-sensitive hashing (LSH) into the fuzzy clustering approach F*k*-centers [156] to reduce dimensions. This process involves applying LSH to predict initial fuzzy clusters in a low-dimensional space. The LSHF*k*-centers algorithm is an extension of LSH-based methods for hard clustering [157]. Despite its effectiveness compared to benchmark algorithms, the computational time of LSHF*k*-centers remains higher than that of its original method. Moreover, it is even more time-consuming than other membership chromosome-based techniques, such as the MaOfCentroids algorithm. Hence, alternative measures other than distance learning dissimilarity for categorical data (DILCA), such as context-based dissimilarity measures, can be explored. Additionally, to enhance locality-sensitive factors, utilizing properties of multi-attributes as the LSH hash function is recommended.

(3) Rough-set-based clustering

Table 12 presents an overview of the different algorithms for rough-set-based clustering.

- RST based on the *K*-modes algorithm

Fuzzy set theory and rough set theory (RST) represent two common approaches for handling uncertainty in data. However, they employ distinct techniques. While fuzzy set theory assigns membership degrees within the range of 0 to 1, with 0 indicating no membership and 1 indicating full membership, RST tackles uncertainty by discerning lower and upper approximations.

In their work to enhance the *K*-modes algorithm, Suri and Murty proposed the rough *K*-modes (R*K*Modes) algorithm [85], integrating lower and upper approximations from rough sets. This method, employing Cao's initialization technique [115] for cluster initialization, iteratively maximizes the modes' density until convergence, thereby introducing an effective approach to outlier detection within the *K*-modes framework.

Another algorithm, known as the density rough *K*-modes (DR*k*-M) algorithm [70–72], has been proposed to address the issue of random selection during the update of modes in the *K*-Modes algorithm. The DR*k*-M algorithm calculates the density of the modes and subsequently applies RST to select the most suitable modes based on the concepts of lower and upper approximations in RST.

Table 12. Rough-set-based clustering.

Authors (Year)	Algorithms	Methods	Comparisons
Ammar et al. (2015)	semantically segmented clustering based on possibilistic and rough set theories [167]	*K*-modes algorithm based on possibility and rough set theories (KM-PR) with semantic interpretations as a discretization method	n/a
Park & Choi (2015)	ITDR [62]	RST integrated with possibility based on information-theoretic attribute dependencies to handle uncertainty in values of attributes and uncertain clusters	*K*-Means, FKM, Fuzzy Centroids [150], SDR [168], SSDR [169], MMR [99]
Suri et al. (2016)	R*K*Modes [85]	*K*-modes algorithm based on RST for outlier detection	*K*-Modes, MMR [99], MTMDP [38]
Yanto et al. (2016)	MF*k*-PIND [42]	fuzzy *k*-Partition based on indiscernibility relation	Fuzzy Centroids [150] and Fuzzy *k*-Partition [170]
Xu et al. (2019)	FRC [49]	*K*-modes algorithm based on RST with the information granularity and dimension reduction method	Cao [115], WKModes [114], *K*-modes
Saha et al. (2019)	SARFKMd, GARFKMd, IRFKMd-RF [50]	the rough fuzzy *K*-modes (RFKMd) with random forest and the metaheuristic methods (simulated annealing, GA)	ccdByEnsemble [171], G-ANMI [101], MMR [99], Tabu Search based FKM [172], AL [173], FKM, RFKMd [174], Rough *K*-medoids [175], *K*-medoids [118], *K*-modes
Naouali et al. (2020)	DR*K*-M [72]	RST uses the density to update the modes	*K*-modes, original weighted *K*-modes [176], original Ng's *K*-modes [177], improved weighted *K*-modes [39], improved Huang's *K*-modes [39], improved Ng's *K*-modes [39]
Salem et al. (2021)	DR*K*-M [70]	RST uses the density to update the modes	*K*-modes, Ng's *K*-modes [143], Cao [115]
Salem et al. (2021)	DR*K*-M [71]	RST uses the density to update the modes	*K*-modes, Ng's *K*-modes [143], Cao [115], the improved Huang's *K*-modes, the Weighted *K*-modes [39], improved Ng's *K*-modes, Bai [178], Khan [121], FKM
Uddin et al. (2021)	MVA [84]	the concept of a number of automated clusters (NoACs) with a rough value set	MDA [103], MSA [179], ITDR [62]

Moreover, Ammar et al. integrate possibility theory with RST, aiming to manage uncertainty in attribute values by utilizing possibility degrees and uncertain clusters through possibilistic membership degrees. This approach extends their prior work [180] by employing a discretization method to convert numeric values into semantically more meaningful linguistic variables with possibilistic memberships based on the *K*-modes algorithm [167].

- RST based on information theory

Park and Choi introduced the information-theoretic dependency roughness (ITDR) [62]. This algorithm concentrates on the dependencies of information-theoretic attributes, employing rough attribute dependencies in categorical-valued information systems to select clustering attributes based on their rough entropy values. Furthermore, ITDR employs a divide-and-conquer approach for object splitting and utilizes the mean degree of rough entropy to select the partition attribute. However, the ITDR algorithm still encounters challenges associated with entropy roughness in identifying the clustering attribute.

Therefore, Uddin et al. introduced the maximum value attribute (MVA) algorithm [84], which integrates the concept of the number of automated clusters (NoACs) to improve cluster purity while reducing complexity compared to other existing rough sets-based clustering algorithms. The MVA algorithm, which adopts the principles of RST, contains three main steps: (1) computing the value sets for each attribute, (2) determining the cardinality of each attribute value set, and (3) selecting the clustering attribute based on the maximum cardinality of the value set. By adopting this approach, the MVA algorithm effectively handles the limitations and issues associated with the random selection of clustering attributes, particularly in cases of independence and insignificant data. Comparative evaluations demonstrate that the MVA algorithm outperforms existing rough sets-based clustering algorithms, including the ITDR algorithm.

- RST based on fuzzy *k*-partition algorithm

In addition to FKM, other popular fuzzy clustering methods include fuzzy *k*-partition (F*k*P) [170] and fuzzy centroids [150]. Yanto et al. proposed a modification of F*k*P known as modified F*k*P based on indiscernibility relation (MF*k*-PIND) to address the limitations of F*k*P, such as high computational time and low clustering purity. Unlike F*k*P, which relies on the likelihood function of multivariate multinomial distributions, MF*k*-PIND is based on the indiscernibility relation. Thus, the MF*k*-PIND algorithm outperforms both FkP and Fuzzy Centroids in terms of clustering performance.

- Fuzzy rough clustering

Saha et al. integrate the rough fuzzy *K*-modes (RFKMd) algorithm with metaheuristic methods. Therefore, the resulting algorithms are called SARF*K*Md when RFKMd is integrated with simulated annealing, and GARFKMd when integrated with genetic algorithms. Both are referred to as SARFKMd-RF and RFKMd-RF when combined with random forest (RF). These algorithms are based on a generalized approach termed integrated rough fuzzy clustering using random forest (IRFKMd-RF) [50]. The utilization of metaheuristic methods aims to optimize the initial cluster modes, addressing the issue of indiscernibility and vagueness inherent in RFKMd, which often leads to local optima.

Furthermore, random forest trains the central points to classify peripheral points and their subsets effectively, including semi-best and pure peripheral points. The roughness measure is then utilized to select the best central points among the three algorithms, aiming to improve clustering performance.

Moreover, Xu et al. introduced a fuzzy rough clustering (FRC) algorithm [49] based on RST, combining information granularity and dimension reduction. FRC employs a weighted distance metric to measure dissimilarity in categorical datasets by converting them into numerical datasets. This conversion enables the utilization of manifold learning techniques to reduce the dimensionality of data points, resulting in decreased complexity compared to using the rough set algorithm directly.

3.2.3. Distance Function

Table 13 presents a summary of dissimilarity functions proposed between 2014 and 2023.

Table 13. Dissimilarity function.

Authors (Year)	Algorithms	Measurement-Based	Comparisons
Lee & Lee (2014)	CATCH [181]	value difference (VD) and value distribution-oriented dimensional weight (VOW) to cluster the high-dimensional multi-valued data	Jaccard coefficient, which is embedded with the *K*-modes algorithm
Chen et al. (2015)	Subspace clustering of categories (SCC) [61]	probabilistic distance function based on kernel density estimation	non-mode clustering (KR) [126], WKM [130], mode-frequency-based (MWKM) [131], complement-entropy-based (CWKM) [114]
Chen (2016)	KPC [77]	a probability-based learning framework with a kernel smoothing method to optimize the attribute weights	*K*-modes, DWKM [130], MWKM [131], CWKM, and EBC [182]
Qian et al. (2016)	SBC [59]	space structure-based representation scheme	*K*-modes, Chan [130], Mkm-nof, Mkm-ndm [183]
Jia et al. (2016)	Frequency probability-based distance measure (FPDM) [58]	frequency probability and co-occurrence probability	Hamming distance (HD) [184], Ahmad's distance [185]
Jiang et al. (2016)	*k*-MODET (Ini_Distance, Ini_Entropy) [37]	traditional distance-based outlier detection technique [186], partition entropy-based outlier detection technique	Khan [121], Cao [125], Wu [122], and the random initialization method embedded with the *K*-modes algorithm.
Amiri et al. (2018)	EN-KM, EN-MBC, EN-SL, EN-AL, EN-CL [94]	ensembled dissimilarity based on the hierarchical method and Hamming distance	*K*-modes, DBSCAN [187], ROCK [107], MBC [188], ensembled version of *K*-modes and MBC using dissimilarity matrix D (EN-KM and EN-MBC), agglomerative (SL, AL, CL, EN-SL, EN-AL, EN-CL)
Jian et al. (2018)	A coupled metric similarity (CMS) measure [88]	the intra-attribute similarity (frequency-based) integrated with the inter-attribute measure (correlation-based)	ALGO [189], coupled object similarity (COS) [190,191], distance matrix (DM) [58], occurrence frequency-based measure (OF) [192], and HD [193] embedded with spectral clustering [194] and *K*-modes algorithms.
Chen & Yin (2018)	CWC [87]	a non-center-based algorithm based on weighted similarity.	OF [195], Goodall3 [192,196], and MSFM measures embedded with *K*-modes [197], KPC [77], entropy-weighting *K*-modes (CEWKM) [114], and the MWKM algorithm [131]
Sulc & Rezankova (2019)	VE, VM [48]	a relative frequency-based where the VE measure uses the entropy while the VM measure uses the Gini coefficient.	ES [198], G1, G2, G3, G4, LIN1 [192], MZ [199], OF, IOF [195], LIN [200], and simple matching [201] embedded with three linkage methods of hierarchical cluster analysis

Table 13. *Cont.*

Authors (Year)	Algorithms	Measurement-Based	Comparisons
Ye et al. (2019)	Heterogeneous Graph-based Similarity measure (HGS) [52]	a heterogeneous weighted graph combining the content-based and structural-based similarity measures	HD [143], OF [18], Lin [200], ALGO [189], and CMS [88] embedded with spectral clustering (SC) and *K*-modes algorithm
Zhang et al. (2020)	EBDM [75]	entropy-based distance metric with a weighting scheme for the mixed-categorical attributes	Hamming Distance, Ahmad's distance [185], ABDM [189], CBDM [202,203], CDDM [58] embedded with *K*-modes, WKM [176], entropy weighting (EW) *K*-means [204], WOC and EBC [182]
Yuan et al. (2020)	mixed-type dissimilarity measure [78]	the idea of mining ordinal information and the rough set theory for the mixed-categorical attributes	Huang, Cao [125], SBC [59], and CMS [88] embedded with *K*-modes
Zheng et al. (2020)	SBC-C [91]	space structure-based representation scheme	SBC, SC [132], *K*-modes, One-Hot Encoding
Rios et al. (2021)	learning-based dissimilarity [83]	a classification ensemble to compute a confusion matrix for the attribute	Eskin [198], Lin, OF, IOF, Goodall, Gambaryan, Euclidean, and Manhattan embedded with *K*-means++
Zhang & Cheung (2022)	UDM [73]	entropy-based distance metric using the weights attributes	Distance measures: HD [184], Goodall [196], Lim [200], context-based distance metric (CBDM) [203], FPDM [58] and EBDM [75] are embedded into *K*-modes, entropy-based categorical data clustering (ECC) [182], the representative attribute weighting *K*-modes (WKM) [130], mixed attribute WKM (MWKM) [131], and SCC [61], and WOC [205]
Zhang & Cheung (2022)	HD-NDW [74]	an automatic distance weighting mechanism based on the intrinsic connection of ordinal and nominal attributes	HD, Lin, CBDM, FPDM, EBDM, CMS embedded with *K*-modes, ECC, WKM, MWKM, and attribute Weighting, WOC, SBC, Coupled Data Embedding-based clustering (CDE) [206], UNsupervised heTerogeneous couplIng IEarning-based clustering (UNTIE) [207], Distance Learning-based Clustering (DLC) [208]
Kar et al. (2023)	an entropy-based dissimilarity measure [86]	Bolzmann's entropy [209]	Distance Measure with Entropy (DME) [210], HD [211], Weighted Similarity Measure (WSM) [109], FPDM, Gambaryan [212], Burnaby [213], embedded with *K*-modes, weighted *K*-modes [176] and Density Peak Clustering for Mixed Data (DPC-MD) algorithms [210]

Table 13. *Cont.*

Authors (Year)	Algorithms	Measurement-Based	Comparisons
Zhang et al. (2023)	MAP, BFKMG [89]	Bayesian dissimilarity measure to measure the dissimilarity, Kullback–Leibler (KL) divergence-based regularization to find the patterns in datasets	Cao [125], FKMFC [150], KL-FCM-GM [214], MWK-DC [215], SBC-C, CFE [216], UDM

- Distance metric based on the VD and VOW

In 2014, Lee and Lee introduced CATCH [181], a categorical data dissimilarity measure designed to cluster high-dimensional multi-valued data effectively. CATCH distinguishes the level of difference between categorical values using the value difference (VD). It incorporates the implicit influence of each attribute on constructing a particular cluster through value distribution-oriented dimensional weight (VOW).

- Kernel-based method

Chen et al. and Chen proposed two algorithms for clustering high-dimensional data into subspaces: the subspace clustering of categories (SCC) algorithm [61] and the *K*-means-type projective clustering of the categorical data (KPC) algorithm [77].

The SCC algorithm is a partition-based clustering approach that utilizes kernel density estimation (KDE) to assign a weight to each attribute, reflecting the smoothed dispersion of categories within a cluster. Furthermore, it employs a probabilistic distance function to measure dissimilarity between data objects and defines a cluster validity index for estimating the number of clusters. Further improvement involves assigning individual weighting exponents to each cluster and adaptively estimating parameters. Additionally, the method can be extended to general kernel functions and tested across various kernels. Similarly, the KPC algorithm uses a probability-based learning framework, leveraging KDE to optimize both attribute weights and cluster centers.

The clustering with weighted categories (CWC) algorithm also conducts subspace clustering. Unlike the KPC algorithm, CWC is non-center-based. CWC performs better on most datasets due to its adaptive learning of distances based on category heterogeneity instead of relying on the independence assumption for computing object-to-cluster distances. However, KPC typically requires less computational time compared to CWC.

- Space structure-based method

Qian et al. [59] introduced a novel data representation scheme that maps categorical objects into Euclidean space, where each object corresponds to a single coordinate. This scheme forms the basis of the space structure-based clustering (SBC) framework. For instance, SBC utilizes Euclidean and cosine distances during experimentation, comparing their performance with various *K*-modes-type algorithms.

However, due to the time-consuming computation of similarity matrices for large datasets and the increase in dimensionality based on the number of datasets, the SBC algorithm required heavy memory loads and high computational complexity. To address these challenges, Zheng et al. proposed the space SBC algorithm with pre-clustering (SBC-C) [91]. SBC-C tackles the limitations of the SBC algorithm by employing two strategies: selecting an appropriate reference set and combining the *K*-means algorithm with the proposed representation. This strategy differs from SBC, which directly applies the *K*-means algorithm to the entire representation.

- Learning-based dissimilarity method

Rios [83] introduced a learning-based dissimilarity approach that focuses on capturing per-attribute object similarity rather than relying on attribute interdependence. This dissimilarity measure aims to identify correlations between values of categorical attributes

through ensemble classification. If such correlations indicate a similarity relation, they assist in determining the appropriate cluster for each object.

An advantage of the learning-based dissimilarity approach is its ability to predict the values of a target attribute. Consequently, this measure can be applied effectively in classification tasks.

- Coupled similarity learning method

Jian et al. proposed another measure known as coupled metric similarity (CMS) [88], which is designed to assess the intrinsic similarity of categorical data, particularly data that is not independent and identically distributed (non-IID). CMS is capable of flexibly capturing both intra-attribute and inter-attribute couplings, as well as value-to-attribute-to-object hierarchical couplings to measure object similarity.

In scalability testing, CMS demonstrated significantly faster and superior capability in capturing couplings compared to other similarity measures. Furthermore, CMS can be integrated with feature selection or weighting techniques to increase effectiveness and efficiency. Additionally, CMS has the potential to be extended for handling heterogeneous data, designing data structures for scalable clustering, and automatically determining the strength of couplings in the data.

- The mixed categorical attributes (nominal and ordinal) method

The HD-NDW algorithm [74], or homogeneous distance–novel distance weighting, is a clustering algorithm that incorporates HD intra-attribute information, focusing on the intrinsic connection between ordinal and nominal attributes. At the same time, the NDW calculates the weights of intra-attribute distances defined by HD to achieve optimal clustering results.

Furthermore, the authors of HD-NDW also introduced two additional methods personalized for mixed categorical attributes: the unified distance metric (UDM) [73] and the entropy-based distance metric (EBDM) [75]. Both UDM and EBDM are centered around information-theoretic principles, utilizing entropy-based distance metrics.

The EBDM unifies distance measurement by incorporating order information from ordinal attributes and statistical information from nominal attributes. Additionally, a unified attribute weighting scheme is introduced to differentiate attribute contributions. However, clustering performance can be improved if EBDM incorporates valuable information from other attributes. Thus, Zhang and Cheung proposed the UDM, which considers intra-attribute and inter-attribute statistical information in distance measurement. Despite its effectiveness, UDM falls short of algorithms like MWKM and SCC, which are specifically designed for nominal data.

Similarly, the dissimilarity measure introduced by Yuan et al. [78] is designed for both ordinal and nominal attributes. This method offers a dissimilarity measure for ordinal attributes, quantifying the degree of ordering based on rough set theory. Comparative analysis against previous algorithms, such as SBC [59] and CMS [88], demonstrates superior performance in measuring ordinal attributes.

- Distance metric based on Graph

Another approach to measuring dissimilarity is the heterogeneous graph-based similarity (HGS) proposed by Ye et al. [52]. First, a heterogeneous weighted graph is constructed to capture latent relationships among attributes. Additionally, HGS considers both the occurrence and co-occurrence relationships between objects and attributes. Leveraging this concept, the similarity measure for objects and attribute values, including their structures, is iteratively calculated until convergence.

- Information-theoretic based approach

Kar et al. [86] introduced an entropy-based dissimilarity metric inspired by Boltzmann's principles of counting microstates to cluster diverse datasets. This dissimilarity measure calculates the entropy of each attribute, followed by determining the weight of each attribute to indicate its significance in the dataset.

Similarly, Jiang et al. [37] employ an information-theoretic-based approach to select initial clusters. They utilize a weighted matching distance metric named initialization *K*-modes using outlier detection (k-MODET). This approach integrates traditional distance-based outlier detection techniques (ini_distance) with partition entropy-based outlier detection techniques (ini_entropy).

- Frequency-based approach

The most cited article in this study, as mentioned in Section 3.1, is by Jia et al. [58]. They proposed a novel distance metric to measure the distance between categorical data. This metric is based on frequency probability, enabling the measurement of the distance of each attribute value in the entire dataset. Moreover, they introduced a dynamic weighting scheme to adjust the contribution of each attribute distance to the overall object distance. The proposed distance metric encompasses three cases: (1) frequency probability-based, (2) adjusted distance metric with dynamic attribute weight without considering the relationship between attributes, and (3) the complete distance metric. Lastly, considering that some attributes are interdependent, the degree of dependency between each pair is calculated using frequency probability and frequently co-occurring items.

Sulc and Rezankova [48] also employed a frequency distribution of categories to address attributes with more than two categories. Their proposed variability-based similarity measures include the variable entropy (VE) and the variable mutability (VM) measure, integrated with three hierarchical cluster analysis linkage methods.

- Ensemble dissimilarity based on hierarchical clustering

Amiri et al. [94] introduced another dissimilarity measure based on hierarchical clustering. Their approach focuses on ensembled dissimilarity designed for datasets with low, high, and varying dimensions. For high-dimensional data, categorical vectors are separated into equal and unequal lengths by including an additional layer of assembly. Alignment procedures are then employed to standardize the unequal categorical vectors. The results demonstrate improved performance of the ensembled clustering method under average linkage (AL) or complete linkage (CL). Currently, due to the absence of clustering methods for unequal-length categorical vectors, the proposed approach can only be compared with the output of phylogenetic trees.

- Bayesian dissimilarity and KL divergence approach

In 2023, a novel fuzzy clustering objective function was introduced, leveraging the concept of approximating the maximum a posteriori (MAP) and employing a Bayesian dissimilarity measure [89]. Moreover, to increase clustering performance, the objective function includes Kullback–Leibler divergence-based graph regularization to identify patterns within datasets.

3.2.4. Weighting Method

The summary of the weighting method is provided in Table 14.

- Automatic feature weight

WF*K*-modes [80], proposed by Sara and Das in 2015, is an automated feature weight learning method designed to adjust feature weights based on their contributions to clustering adaptively. It aims to minimize the objective function and determine cluster membership within the FKM algorithm [217]. Experimental results indicate that this algorithm performs effectively, especially in datasets containing noise features. Despite trying to modify the algorithm for scenarios with an unknown number of clusters, it still requires setting threshold values for the maximum and minimum number of clusters. Thus, further studies can explore the effectiveness of different cluster validity indices and their relationship with the weight vector. Additionally, parameters such as "β" associated with attribute weight and objective function minimization need optimization. Furthermore, extending performance evaluation to larger datasets would be beneficial.

Table 14. Weighting method.

Authors (Year)	Algorithms	Comparisons
Saha & Das (2015)	WF*K*-modes [80]	n/a
Kim (2017)	attribute weighting method based on within-cluster and between-cluster impurity measures [95]	*K*-modes, FKM
Peng & Liu (2019)	weighting method combined with the distance and density measures to select the cluster centers based on a rough set and information theory [51]	Random method, Khan [121], Cao [125], Wu [122]
Oskouei et al. (2021)	FKMAWCW [79]	Initialization sensitivity reduction methods: Khan [121], Cao [125], Wu [122], *k*-MODET [37], Peng [51], Mod-2 [127], Mod-3 [128], and Attribute-weighted method: IWFKM [55], EWKM [114], Saha [80], SBC [59], Chan [130], Jia [205]

Additionally, Oskouei et al. [79] explored automated attribute weighting, extending the work of [218], which employed cluster weighting to select initial centers in the FCM algorithm. However, since FCM primarily handles numerical attributes, their proposed method, the categorical fuzzy *K*-modes clustering with automated attribute-weight and cluster-weight learning (FKMAWCW) algorithm, is implemented for categorical attributes. This algorithm uses a local attribute weighting mechanism to appropriately weigh attributes within each cluster and a cluster weighting mechanism to address initialization sensitivity. Furthermore, to mitigate noise sensitivity, they introduce a novel distance function combining frequency probability-based distance [58] and non-Euclidean distance [219]. Exploring the suitability of the FKMAWCW algorithm for clustering mixed data, especially considering its emphasis on categorical data, would be valuable. Moreover, future studies can explore the automatic determination of the number of clusters during the clustering process.

- Information-theoretic approach

Kim [95] introduced a novel attribute weighting approach for the *K*-modes and FKM algorithms based on within-cluster and between-cluster impurity measures to identify attribute relevance in separate clusters. These impurity measures, such as entropy and Gini impurity, assign large weights to variables with lower entropy or Gini impurity. However, the effectiveness of attribute weights depends on the parameter "c", which controls the balance between within-cluster and between-cluster information. Determining the optimal value for "c" relies on general guidelines and requires further investigation. Furthermore, the proposed method can be expanded to accommodate numerical and mixed attributes by employing inhomogeneous measures for numerical features.

Peng and Liu [51] aim to improve the cluster center during the initialization phase of the *K*-modes algorithm by employing an attribute-weighted distance metric and weighted average density rather than relying solely on the simple matching distance metric. This approach helps prevent the possibility of outliers becoming cluster centers or multiple cluster centers converging around a single center. Furthermore, this approach can broaden its scope in future studies by employing feature selection techniques to identify significant attributes for distance measurement between instances during cluster center initialization.

3.2.5. Validity Function

Table 15 shows the summary of cluster validity.

Table 15. Cluster validity.

Authors (Year)	Function	Comparisons
Bai & Liang (2014)	BCI*k*-M [39]	Ng's *K*-modes [143,220], *K*-modes [15,221], WKM [176]
Bai & Liang (2015)	generalized validity function [65]	*K*-modes, CU [222], IE [102]
Gao & Wu (2019)	IDC, CUBOS [57]	CCI [223], CDCS [224], IE, CU, NCC [225]

All the validity functions in this study focus on internal validity functions. In 2014, Bai and Liang [39] improved the *K*-modes algorithm by optimizing its objective function to incorporate both between-cluster separation and within-cluster compactness. Their proposed algorithm, named between-cluster information *K*-modes (BCI*k*-M), demonstrated improved effectiveness compared to traditional FKM algorithms. The integration of between-cluster information with the FKM algorithm, as verified in [178], shows its superior effectiveness. Furthermore, this study enhanced several original *K*-modes algorithms, including Ng's *K*-modes [143,220], Huang's *K*-modes algorithm [15,221], and the weighted *K*-modes algorithm (WKM) [176], by including both types of information. Despite the increased computational time required by the improved *K*-modes algorithm in scalability tests, the increase rate remains linear, guaranteeing its effectiveness and scalability.

In 2019, Bai and Liang [57] introduced a study focusing on the generalized validity function. Initially, they examined three existing internal validity functions: *K*-modes [221], category utility function (CU) [222], and information entropy function (IE) [102]. As these functions solely relied on within-cluster information, the study aimed to investigate the impact of including between-cluster information on performance. The experimental results demonstrated that these three validity functions effectively evaluated clustering results even without utilizing between-cluster information. Additionally, the study proposed normalizations for these internal validity functions and found that normalization increases performance.

Gao and Wu [57] conducted a comprehensive review of existing functions of internal validity indices and, based on that, proposed the categorical data cluster utility based on silhouette (CUBOS). The CUBOS method combines the Silhouette index with an improved distance metric for categorical data (IDC). It considers the relationship between different attribute values and the detailed distribution information among data objects. IDC represents a novel improvement measure inspired by category distance [87]. Furthermore, the CUBOS framework facilitates a more detailed distribution of information within clustering results.

In addition to the hierarchical and partition clustering methods, as well as the dissimilarity functions, weighting methods, and cluster validity measures outlined in Tables 8–15, it is relevant to highlight the datasets utilized in these studies. Each study employs different datasets depending on its objectives, although the datasets may not always be categorized based on their scalability or dimensions. The following section combines several frequently used datasets with several validity functions.

3.2.6. Datasets

The total number of datasets used in all articles is 51, as outlined in Table 16. Five of these datasets are extensively featured in over 30 articles. These datasets include breast cancer Wisconsin (original), congressional votes, mushroom, soybean small, and zoo. The summary aligns with the original dataset specifications from the UCI Repository [12], including the information on the number of records (#rec), attributes (#attr), and clusters (#clus).

Table 16. Datasets.

Datasets	#rec	#attr	#clus	n	Datasets	#rec	#attr	#clus	n
Adult + Stretch	48,842 *	14 *	2	3	HIV-1 protease cleavage	6590	9 *	2	1
Arrhythmia	452	279	16	1	Horse Colic	368 *	24 *	2	1
Audiology	226	69	24	2	Letter Recognition (E, F)	1543	16	2	4
Australian Credit Approval	690	14	2	1	Lung Cancer	286	9	2	17
Balance	625	4	3	12	Lymphography	148 *	18	4 *	17
Ballonn	20	4	2	6	Mammographic Masses	961 *	4	2	3
Breast Cancer Wisconsin (Original)	699 *	9	2	38	Microsoft Web	37,711	294	-	3
Car Evaluation	1728	6 *	4 *	1	Monk	432	6	2	5
Cervical Cancer	858	32 *	4	1	Mushroom	8124	22 *	2	38
Chess	3196	36	2	14	Nursery	12,960	8 *	3 *	15
Chess (Big)	28,056	6	18	1	Page Blocks	5473	10	5	1
Congressional Votes	435 *	16	2	37	Primary Tumor	339 *	17 *	21	7
Connect-4	67,557	42	3	4	Post-Operative Patient	90 *	8	3	2
Contraceptive Method Choice	1473	10	3	1	Shuttle Landing Control	15	6	2	2
Credit Approval	690 *	15 *	2	10	Solar Flare	1066	10 *	6	9
Dermatology	366	34 *	3	16	Soybean Large	307 *	35	19 *	6
DNA Splice	3190	60	3	10	Soybean Small	47	35 *	4	43
DNA Promoter	106	57	2	14	Spect Heart	267	22	2	10
Drug Consumption	1885	6 *	7	1	Sponge	75	45	12	1
Fitting Contact Lenses	24	4	3	8	Student	300	32	3	1
Flag	194	30	-	1	Thoracic	470	16	2	1
Germany	1000 *	20	2	1	Tic-Tac-Toe	958	9	2	17
Hayes-Roth	132	4	3	17	Train	10	32	2	1
HCC survival	165	49 *	2	1	Optical Recognition of Handwritten Digits	5620 *	64	10	1
Heart Disease	303	8	2	8	Zoo	101	16 *	7	40
Hepatitis	155 *	19 *	2	3					

* Each algorithm employs a distinct number.

3.2.7. Performance Evaluation

Table 17 presents the performance evaluation methods employed in the 64 articles. Among these, 20 internal and 13 external validity functions were utilized. Notably, the accuracy, adjusted rand index, and normalized mutual information were employed in oven 20 articles. Furthermore, Table 18 illustrates the most frequently used validity functions corresponding to the most common datasets.

Table 17. Performance evaluation.

No.	Internal Validity Functions	n	No.	External Validity Functions	n
1.	Silhouette coefficient	4	1.	Accuracy (AC)	42
2.	Davies–Bouldin index (DBI)	5	2.	Adjusted rand index (ARI)	32
3.	Category utility function (CU)	4	3.	Random index (RI)	5
4.	Dunn	2	4.	Normalized mutual information (NMI)	21
5.	Calinski–Harabasz index (CH)	2	5.	Purity	12
6.	New Condorcet criteria (NCC)	1	6.	Entropy	8
7.	Compactness	1	7.	Precision (PE)	10
8.	Separation	1	8.	Recall (RE)	10
9.	Fuzzy silhouette coefficient (FSI)	1	9.	F-measure	9
10.	Multivariate FSI (MFSI)	1	10.	Jaccard coefficient	2
11.	Sum of square error (SSE)	1	11.	Micro-p	1
12.	Pseudo F index based on the mutability (PSFM)	1	12.	Fowlkes and mallows index (FM)	1
13.	Pseudo F index based on the entropy (PSFE)	1	13.	Roughness measure	2
14.	Partition entropy coefficient (PE)	1			
15.	Partition coefficient (PC)	1			
16.	Cluster cardinality index (CCI)	1			
17.	Categorical data clustering with subjective factors (CDCS)	1			
18.	Information entropy (IE)	1			
19.	Czekanowski–Dice index (CDI)	1			
20.	Kulczynski index	1			

Table 18. Validity function summary.

Validity Index	#Articles				
	Breast Cancer Wisconsin (Original)	Congressional Votes	Mushroom	Soybean Small	Zoo
AC	19	20	20	22	21
ARI	16	17	15	21	21
NMI	12	10	8	12	11

In Tables 19–21, algorithms are classified based on their types to illustrate the best algorithm with the best result, aligned with the most frequent datasets and validity indexes employed in the articles. The validity indexes include accuracy, ARI, and NMI. As several articles propose more than one algorithm, the summary only presents the best algorithm along with its corresponding results.

Despite the limited number of datasets and validity indexes used in performance evaluation, valuable insights can still be provided. In this study, articles employing rough-set-based clustering do not utilize the ARI, and none of the weighting methods employ the NMI as the validity index.

Many algorithms that use the soybean dataset have the highest value for all the validity indexes. However, for the mushroom dataset, only the EN-CL algorithm can achieve an accuracy of 100%, and the ARI value is equal to one. It shows that compared to other methods, the EN-CL, which is the ensembled dissimilarity, can achieve better results, especially for high-dimensional and scalable datasets. Moreover, all weighting method algorithms used in the soybean small dataset have the highest accuracy, and one of the weighting methods, the FKMAWCW, also has the highest ARI value. The cluster validity algorithm, such as CUBOS [57], also has the highest value for all validity indexes using the same dataset.

Table 19. Accuracy results.

Algorithm *	Breast Cancer Wisconsin (Original)	Congressional Votes	Mushroom	Soybean Small	Zoo
Hierarchical Clustering					
MNIG [47]	92.7	87.4	84.8	97.9	93.1
MGR [41]	88.4	82.8	67.7	-	93.1
HPCCD [96]	-	92.18	86.41	100	96.04
P-ROCK [106]	-	79.77	-	-	-
Partition Clustering: Hard Clustering					
MF*k*-M [64]	-	-	45	-	-
SKSCC [76]	96.59	87.34	81.94	90.85	80.43
DKBKM-Max [111]	79.1	82.19	81.7	-	74.5
MOCSG [81]	89.1	-	-	100	83.2
k-PbC [98]	96.14	88.05	88.61	100	89.11
Partition Clustering: Fuzzy Clustering					
GAFKM [2]	-	86.6	-	-	-
GIWFKM [55]	69.2	91	93.2	98.5	92.7
MaOFcentroids [63]	-	88.1	88.5	100	91
SCA-PFKM [69]	94.07	86.44	88.95	100	-
KIWFKM-DCP [93]	70.28	-	88.31	98.72	90.1
Partition Clustering: Rough-set-based Clustering					
IRFKMd-RF [50]	-	88.79	91.50	99.85	98.38
DR*k*-M [70]	93.29	-	85.91	-	88.56
DR*k*-M [71]	93.29	-	85.91	100	-
MF*k*-PIND [42]	97.17	-	-	100	89.96
MVA [84]	-	-	-	72	82
Distance Function					
Ini_Entropy [37]	93.28	86.9	88.76	100	90.1
SBC [59]	92.93	87.83	-	96.66	-
SCC [61]	97	-	-	-	-
HD-NDW [74]	65.1	87.6	-	84.9	76
EBDM [75]	-	87.1	-	-	-
EN-CL [94]	-	-	100	100	99
mixed-type dissimilarity measure [78]	-	-	-	95.75	-
CDMs [86]	53.29	86.64	85.05	-	75.91
Weighting Method					
weighted attributes [51]	-	86.71	91.85	100	89.33
FKMAWCW [79]	-	89.22	81.82	100	82.18
Cluster Validity					
CUBOS [57]	78.7	87.9	-	100	-
Improved Ng's *k*-modes [39]	87.7	-	83.66	99.79	89
Total	19	20	20	22	21

* Best algorithm.

Table 20. ARI results.

Algorithm *	Breast Cancer Wisconsin (Original)	Congressional Votes	Mushroom	Soybean Small	Zoo
Hierarchical Clustering					
MNIG [47]	0.725	0.556	0.475	0.937	0.945
MTMDP [38]	0.585		0.274	1	0.96
MGR [41]	0.79	0.8	0.65	-	0.96
HPCCD [96]	-	0.7109	0.5302	1	0.963
Partition Clustering: Hard Clustering					
CDC_DR + SE [66]	0.89	-	0.61	0.74	0.64
MOCSG [81]	-	-	-	1	0.851
OTQT [82]	0.67	-	0.61	0.95	0.66
Partition Clustering: Fuzzy Clustering					
AFC-NSPSO [53]	0.713	0.617	0.634	0.958	0.898
PM-FGCA [54]	0.467	0.624	0.468	0.938	0.832
GIWFKM [55]	0.388	0.649	0.703	0.967	0.93
NSGA-FMC [60]	-	0.508	-	0.919	0.8
MaOFcentroids [63]	-	0.578	0.593	1	0.894
EGA-FMC [90]	-	0.79	-	1	0.92
KIWFKM-DCP [93]	0.5022		0.7967	0.9864	0.9457
Distance Function					
SBC [59]	0.7331	0.5715	-	0.94	-
HD-NDW [74]	0.09	0.564	-	0.803	0.721
EBDM [75]	-	0.548	-	-	-
EN-CL [94]	-	-	1	1	0.99
CDMs [86]	0.0019	0.5349	0.4847		0.7195
BFKMG [89]	0.9138	0.6412	0.4958	1	0.9087
Weighting Method					
FWFKM [95]	0.9111	-	-	0.9787	0.877
FKMAWCW [79]	-	0.6137	0.4053	1	0.7806
Cluster Validity					
CUBOS [57]	0.247	0.574	-	1	-
generalized validity function [65]	0.7712	0.5181	0.6059	1	0.644
Total	19	20	20	22	21

* Best algorithm.

Table 21. NMI Results.

Algorithm *	Breast Cancer Wisconsin (Original)	Congressional Votes	Mushroom	Soybean Small	Zoo
Hierarchical Clustering					
MTMDP [38]	0.541	-	0.443	1	0.925
Partition Clustering: Hard Clustering					
CDC_DR + SE [66]	0.8269	-	0.5845	0.8627	0.7777
MF*k*-M [64]	-	-	0.0962	-	-
Partition Clustering: Fuzzy Clustering					
PM-FGCA [54]	0.507	0.532	0.448	0.882	0.775
KIWFKM-DCP [93]	0.0071	-	0.5632	0.9727	0.8298
Distance Function					
HGS [52]	0.316	0.3	-	0.709	0.753
DM3 [58]	0.6917	0.4987	0.3182	0.8991	0.7927
SCC [61]	0.78	-	-	-	-
HD-NDW [74]	0.062	0.489	-	0.897	0.809
EBDM [75]	-	0.483	-	-	-
CMS-enabled *k*-modes [88]	0.595	0.447	-	1	0.842
BFKMG [89]	0.8503	0.5625	0.4845	1	0.8997
Cluster Validity					
CUBOS [57]	0.144	0.51	-	1	-
generalized validity function [65]	0.6534	0.4555	0.5465	1	0.8071
Total	19	20	20	22	21

* Best algorithm.

Many algorithms utilizing the soybean dataset achieve the highest values across all validity indices. However, concerning the mushroom dataset, only the EN-CL algorithm achieves 100% accuracy, with an ARI value of one. These results indicate that EN-CL, an ensembled dissimilarity approach, achieves better results compared to other methods, particularly for high-dimensional and scalable datasets. Moreover, all weighting method algorithms applied to the soybean small dataset achieve the highest accuracy, and one of these methods, FKMAWCW by Oskouei et al. [79], also secures the highest ARI value. Additionally, the cluster validity algorithm CUBOS by Gao and Wu [57] achieves the highest value across all validity indexes using the same dataset.

3.3. Taxonomy

Numerous taxonomies related to clustering are presented, as outlined in Table 1, with most of them addressing numerical and categorical data. However, as shown in Figure 6, this study aims to construct a taxonomy specifically for categorical data clustering. Nonetheless, this task presents challenges due to the varied perspectives and classification approaches found in each study. To the best of our knowledge, no comprehensive taxonomy has been established for categorical data clustering. Therefore, the proposed taxonomy shown in Table 5 assists scholars by providing a simple yet comprehensive classification that covers all relevant topics in categorical data clustering. This taxonomy is based on mapping domain research in categorical data. However, it has some limitations. For instance, it only covers nominal, ordinal, and mixed data types, excluding sequential categorical data such as DNA.

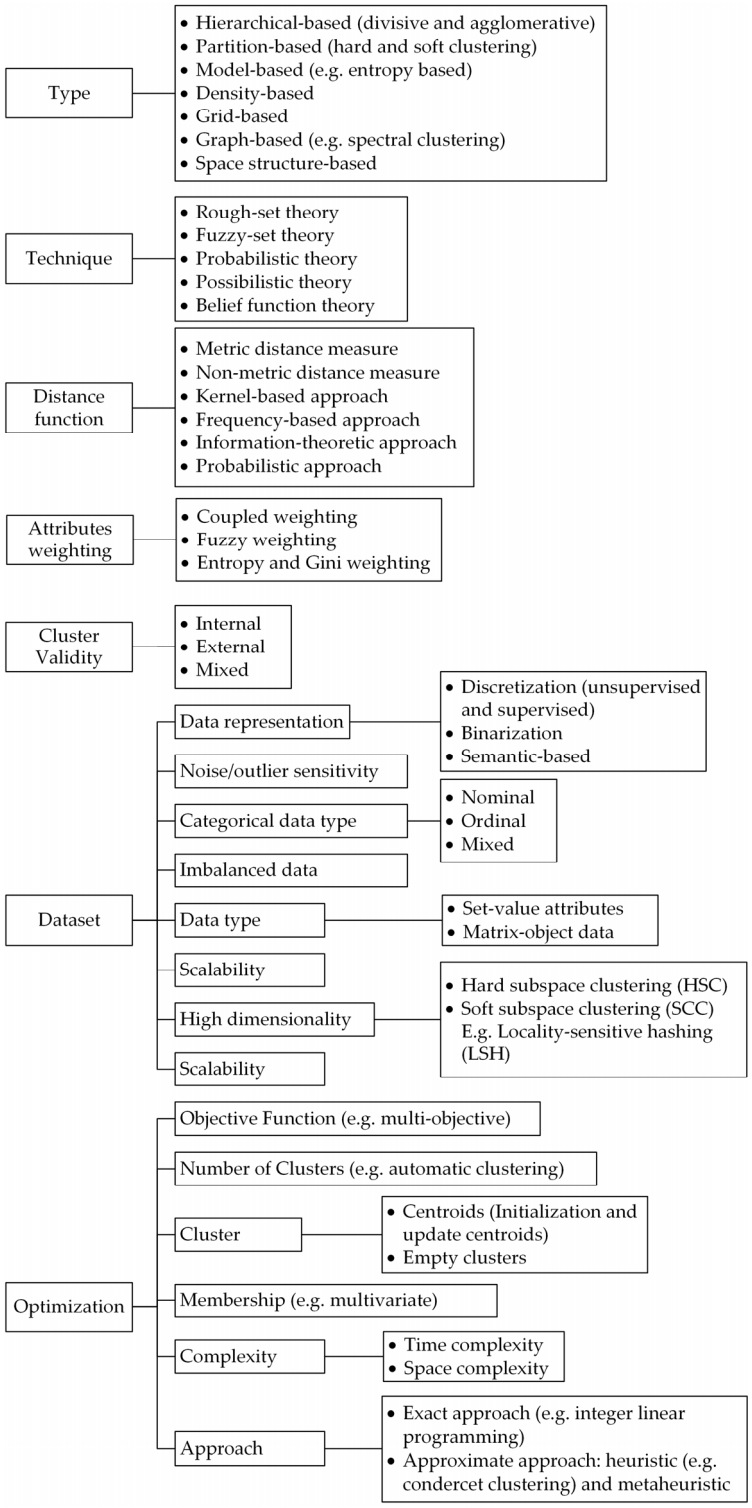

Figure 6. Taxonomy of clustering categorical data.

First, this study adopts the taxonomy [16], which classifies clustering types into hierarchical and partition-based. Hierarchical clustering is further divided into divisive and agglomerative, with the agglomerative approach consisting of single, average, and complete links. Partition-based clustering is divided into hard and soft clustering based on membership degree. In hard clustering, the data points belong to only one cluster, whereas in fuzzy partitioning, the data points can belong to multiple clusters based on their membership degree. Graph clustering is considered a separate category instead of

part of partition-based clustering. The levels of graph clustering are similar to partition-based clustering, alongside other types, including model-based, density-based, grid-based, and space-structure-based. However, space-structure-based clustering may overlap with grid-based or density-based methods since spatial data can be clustered based on density or divided into a grid. Hence, due to the growing research in this area, this study treats space-structure-based clustering as an independent category.

Furthermore, clustering techniques are classified based on background theory, including rough-set, fuzzy-set, probabilistic, possibilistic, and belief functions. These techniques can be applied in hierarchical, partition-based, or other clustering methods as they primarily aim to handle uncertainty and address challenges posed by traditional clustering algorithms. While some references, such as Naouali [17], place hard, fuzzy, and rough-set-based clustering at the same level, with probabilistic and possibilistic methods considered part of a fuzzy theory, this taxonomy assumes all these theories as equal under the category "clustering techniques," aiming to cover the belief functions theory.

Additionally, distance functions and attribute weighting can be integrated with each other. Although only a few attribute weighting methods are listed, including entropy and Gini weighting from the information-theoretic approach, entropy can also be utilized in distance functions. The taxonomy for distance functions is derived from [13,18,192], where the term "distance function" refers to dissimilarity measures. Hence, the taxonomy uses both metric and non-metric distance measures under the term "similarity distance."

Moreover, the concept of context-free and context-relative, as proposed in [18], is applied solely to unsupervised learning, encompassing frequency-based, informatic-theoretic, and probabilistic approaches. On the other hand, this study adds a kernel-based approach since many studies have proposed distance metrics based on the kernel.

Similarly, regarding validation functions, this study found that all methods proposed in the past decade are associated with internal validation. As a result, the validation section remains unchanged, following the previous taxonomy, which consists of internal, external, and mixed validation functions—another updated taxonomy related to datasets and optimization. Instead of combining these two aspects as part of the clustering issue, this study categorizes them based on the source or root cause of the problem. For example, issues such as noise sensitivity, outlier detection, and imbalanced data are caused by the dataset characteristics. Outliers may not always be problematic, as they can be useful depending on the clustering objective. Similarly, addressing high-dimensionality data may involve transformation methods to reduce dimensionality, but this study does not focus on dimensionality reduction or feature selection.

The final aspect relates to optimization. This study divides optimization into several parameters or processes that can be optimized, such as the objective function and the number of clusters. Furthermore, optimization approaches cover both exact and heuristic approaches rather than solely focusing on metaheuristic approaches, as many algorithms still utilize these traditional optimization strategies.

4. Discussion

This study conducts a bibliometric analysis focusing on categorical data clustering, particularly in partition-based clustering. The quantitative synthesis and analysis subsection provides a performance overview of articles and science mapping. A limitation of this study is its dependence solely on articles from the WoS Core Collection from 2014 to 2023. Future studies can expand the scope to include other databases like Scopus or broaden the inclusion criteria. However, the comparison between the 567 and 64 articles over the past decade effectively captures research trends in the field. In the science mapping section, co-word and citation analyses are visualized using VOSviewer. Comparing the top ten most cited articles in Table 3 with the most productive authors in Table 6 reveals interesting insights [59–61,64]. Even though four of the top ten articles are authored by the top ten authors, the most cited article [58] is not written by the most productive author. This comparison and citation analysis provide a deeper understanding of the research trends.

For example, although [58] is cited by 15 of the 64 articles, the second top article is cited by only 9, indicating that citation count alone may not fully capture the significance of a publication. Overall, the citation network provides a comprehensive overview and detailed insights into trends and topics in categorical data clustering, suggesting further analysis related to the relationship between cited articles and authors.

After the quantitative synthesis and analysis, the qualitative synthesis and analysis for the 64 articles are presented. The classification follows a benchmark taxonomy of type, technique, distance function, attribute weighting, validation, dataset, and optimization. The detailed analysis and classification are sequentially presented in Section 3.2.

Specifically, there are five studies related to hierarchical clustering, with four of them based on rough-set theory (MTMDP [38], MGR [41], MNIG [47], and HPCCD [96]), except the P-ROCK [106]. Two studies focus on agglomerative hierarchical clustering, while the remaining three focus on divisive hierarchical clustering.

Related to the hierarchical clustering combined with a rough set, most proposed algorithms show capabilities in handling uncertain and imbalanced datasets, automatically discovering the number of clusters, and clustering high-dimensional datasets. Moreover, these algorithms have improved over the MMR algorithm [99] in terms of increased accuracy and efficiency.

For future research directions, it is suggested that these proposed algorithms can explore the possibility of clustering in scalable and automatic subspace clustering. Extending these algorithms to handle mixed numeric and categorical data can be a promising avenue for further investigation.

Another type of clustering is partition-based clustering. This study proposes an expanded classification of clustering type. Instead of dividing clustering types into hierarchical and partition-based, this taxonomy places all the clustering types at the same level. This includes space structure-based [59,91], which was previously categorized separately. Furthermore, spectral clustering is incorporated into the graph-based category, and entropy-based clustering is now considered part of model-based clustering.

During the period from 2014 to 2023, while the most productive year was 2019, there has been a downward trend since 2021. However, several significant works have emerged, particularly in distance function methods. Other popular research topics include multi-objective optimization clustering based on information-theoretic, kernel-based, and frequency-based approaches. Additionally, many algorithms have been developed to address the challenge of clustering high-dimensional and scalable data.

Many algorithms performed scalability testing, such as those mentioned in references [40,41,56,63,64,67,68,70,96,98,106], aiming to improve clustering methods for high-dimensional data. Notably, algorithms like SCC [61] and SKSCC [76] utilize probabilistic distance functions based on kernel density estimation to increase clustering performance.

Some methods focus on data representation techniques, such as discretization (converting categorical data into numeric values) [64] or representing categorical data as graph structures [66] to reduce time complexity in high-dimensional datasets. Additionally, soft subspace clustering methods like LSHF*k*-centers aim to reduce dimensionality before data processing. However, despite their effectiveness, these algorithms still suffer from high computational time, indicating a need for further research to improve efficiency and reduce time complexity.

On the other hand, several algorithms have been developed based on rough-set theory to address data uncertainty. These algorithms aim to prevent uncertainty associated with attribute values and uncertain clustering outcomes. For example, the RKModes [85] algorithm focuses on outlier detection and sensitivity analysis. While some algorithms are based on the *K*-modes, others, like MF*k*-PIND [42], modify fuzzy *k*-partition (F*k*P) and fuzzy centroids to improve computational efficiency and clustering purity. Additionally, certain algorithms utilize information-theoretic dependencies, such as the widely-used ITDR [62], which employs entropy roughness to identify clustering attributes. However, a challenge arises when clustering attributes possess zero or equal significance values, leading to

random attribute selection. To address this issue, the MVA algorithm [84] was proposed, which overcomes the limitations of ITDR but requires further analysis in combination with other rough purity approaches (RPA).

Several algorithms have been developed to automate clustering by optimizing the number of clusters without requiring a predetermined initialization. One such algorithm is the α-Condorcet algorithm [92], which highlights the practicality of pre-identifying the cluster number in certain real-world scenarios, such as psychometrics. Developed based on a heuristic approach, this algorithm provides valuable insights into cluster number determination. Additionally, metaheuristic approaches have been integrated with clustering algorithms to improve their performance. For instance, fuzzy-based algorithms have been combined with GA, PSO, ABC, and other metaheuristic methods to optimize cluster center initialization [2,55,69]. Furthermore, optimization techniques that combine metaheuristics with multi-objective algorithms have been explored [53,60,90]. However, it is worth noting that while fuzzy-based algorithms utilize various metaheuristics, most multi-objective algorithms primarily rely on GA and PSO. Hence, conducting comparative performance evaluations with other metaheuristic approaches can provide valuable insights, particularly considering the diverse objective functions employed by these algorithms. Additionally, assessing algorithm efficiency in terms of time and space utilization alongside objective function optimization is recommended.

Exploring the characteristics of algorithms capable of handling empty clusters is important, especially considering the commonness of this issue in algorithms like *K*-modes. Many scholars have already used the brute force approaches to address this challenge. In this context, the OTQT algorithm stands out for its adoption of the Hartigan algorithm, a variation of *K*-means, to ensure that clusters remain nonempty during the initialization step. This innovative approach offers a promising solution to prevent the empty cluster problem commonly encountered in categorical data clustering.

Overall, the methods discussed in this study contribute to enhancing the proposed taxonomy. In the future, this taxonomy can serve as a foundation framework for further advancements in clustering algorithms, aligning with the trends identified in the bibliometric analysis. As data complexity continues to increase, there are opportunities to refine existing methods for improvement and innovation. Future research directions may involve integrating clustering methods with deep learning and ensemble techniques and exploring semi-supervised learning approaches capable of clustering mixed labeled and unlabeled data. Furthermore, algorithms can be developed to effectively cluster mixed datasets, thereby improving the overall performance and efficiency of clustering algorithms.

5. Conclusions

The bibliometric analysis conducted between 2014 and 2023, focusing on categorical data clustering and sourced from the WoS Core Collection, identified 64 relevant articles following content screening. Through co-word and citation network analyses, research trends and relationships among publications and clustering topics were presented. Subsequently, a qualitative synthesis and analysis were conducted to explore the details of the studies. The 64 articles were classified according to a previous taxonomy, leading to the development of a new taxonomy based on emerging methods and trends.

Numerous methods were identified to address the limitations of traditional algorithms, particularly in partition-based clustering. These methods include optimization techniques employing metaheuristics and uncertainty methods such as fuzzy and rough-set theory. Various distance functions were proposed to mitigate the shortcomings of simple matching distance, with some considering both within-cluster cohesion and between-cluster separation. Additionally, several attribute weighting methods were introduced to discern the importance of attributes.

This study also synthesized the most commonly used datasets and summarized the performance results. However, it is important to note that no single algorithm can address all clustering challenges, as efficiency depends on factors such as dataset characteristics.

Moreover, this study may not comprehend all issues presented in the articles, and due to the complexity of categorical data clustering, the proposed taxonomy may not cover all methods in detail.

For future works, the majority of studies aim to enhance the existing methods for improved scalability and efficiency while also extending these approaches to accommodate mixed data types beyond categorical datasets. Despite the declining trend observed since 2021 and the numerous algorithms proposed over the past decade for categorical data clustering, certain challenges persist, particularly in addressing issues inherent to traditional algorithms like the *K*-modes-based methods discussed herein. Consequently, it is recommended that modern clustering techniques be explored in future works to tackle these ongoing challenges effectively.

Author Contributions: Conceptualization, M.C. and R.-J.K.; methodology, M.C. and R.-J.K.; validation, M.C. and R.-J.K.; formal analysis, M.C.; investigation, M.C. and R.-J.K.; data curation, M.C.; writing—original draft preparation, M.C.; writing—review and editing, R.-J.K.; visualization, M.C.; supervision, R.-J.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Abdul-Rahman, S.; Arifin, N.F.K.; Hanafiah, M.; Mutalib, S. Customer segmentation and profiling for life insurance using k-modes clustering and decision tree classifier. *Int. J. Adv. Comput. Sc.* **2021**, *12*, 434–444. [CrossRef]
2. Kuo, R.J.; Potti, Y.; Zulvia, F.E. Application of metaheuristic based fuzzy k-modes algorithm to supplier clustering. *Comput. Ind. Eng.* **2018**, *120*, 298–307. [CrossRef]
3. Hendricks, R.; Khasawneh, M. Cluster analysis of categorical variables of parkinson's disease patients. *Brain Sci.* **2021**, *11*, 1290. [CrossRef] [PubMed]
4. Narita, A.; Nagai, M.; Mizuno, S.; Ogishima, S.; Tamiya, G.; Ueki, M.; Sakurai, R.; Makino, S.; Obara, T.; Ishikuro, M.; et al. Clustering by phenotype and genome-wide association study in autism. *Transl. Psychiat* **2020**, *10*, 290. [CrossRef]
5. Farhang, Y. Face extraction from image based on k-means clustering algorithms. *Int. J. Adv. Comput. Sc.* **2017**, *8*, 9. [CrossRef]
6. Huang, H.; Meng, F.Z.; Zhou, S.H.; Jiang, F.; Manogaran, G. Brain image segmentation based on FCM clustering algorithm and rough set. *IEEE Access* **2019**, *7*, 12386–12396. [CrossRef]
7. Wei, P.C.; Zhou, Z.; Li, L.; Jiang, J. Research on face feature extraction based on k-mean algorithm. *Eurasip. J. Image Vide* **2018**, *2018*, 1–9. [CrossRef]
8. Bushel, P.R. Clustering of gene expression data and end-point measurements by simulated annealing. *J. Bioinform. Comput. Biol.* **2009**, *7*, 193–215. [CrossRef]
9. Castro, G.T.; Zárate, L.E.; Nobre, C.N.; Freitas, H.C. A fast parallel k-modes algorithm for clustering nucleotide sequences to predict translation initiation sites. *J. Comput. Biol.* **2019**, *26*, 442–456. [CrossRef]
10. Fonseca, J.R.S. Clustering in the field of social sciences: That is your choice. *Int. J. Soc. Res. Method.* **2013**, *16*, 403–428. [CrossRef]
11. Luo, N.C. Massive data mining algorithm for web text based on clustering algorithm. *J. Adv. Comput. Intell. Inform.* **2019**, *23*, 362–365. [CrossRef]
12. Dua, D.G. UCI Machine Learning Repository. Available online: https://archive.ics.uci.edu/ (accessed on 10 January 2024).
13. Tan, P.-N.; Steinbach, M.S.; Karpatne, A.; Kumar, V. *Introduction to Data Mining*, 2nd ed.; Pearson Education, Inc.: London, UK, 2019.
14. MacQueen, J. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 21 June 1967; pp. 281–297.
15. Huang, Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Min. Knowl. Discov.* **1998**, *2*, 283–304. [CrossRef]
16. Jain, A.K.; Murty, M.N.; Flynn, P.J. Data clustering: A review. *ACM Comput. Surv.* **1999**, *31*, 264–323. [CrossRef]
17. Naouali, S.; Ben Salem, S.; Chtourou, Z. Clustering categorical data: A survey. *Int. J. Inf. Technol. Decis. Mak.* **2020**, *19*, 49–96. [CrossRef]
18. Alamuri, M.; Surampudi, B.R.; Negi, A. A survey of distance/similarity measures for categorical data. In Proceedings of the 2014 International Joint Conference on Neural Networks (IJCNN), Beijing, China, 6–11 July 2014; pp. 1907–1914.
19. Hancer, E.; Karaboga, D. A comprehensive survey of traditional, merge-split and evolutionary approaches proposed for determination of cluster number. *Swarm Evol. Comput.* **2017**, *32*, 49–67. [CrossRef]

20. Alloghani, M.; Al-Jumeily, D.; Mustafina, J.; Hussain, A.; Aljaaf, A.J. A systematic review on supervised and unsupervised machine learning algorithms for data science. In *Supervised and Unsupervised Learning for Data Science*; Unsupervised and Semi-Supervised Learning; Springer: Berlin/Heidelberg, Germany, 2020; pp. 3–21.
21. Awad, F.H.; Hamad, M.M. Big data clustering techniques challenged and perspectives: Review. *Informatica* **2023**, *47*, 6. [CrossRef]
22. Ikotun, A.M.; Absalom, E.E.; Abualigah, L.M.; Abuhaija, B.; Jia, H. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Inf. Sci.* **2022**, *622*, 178–210. [CrossRef]
23. Wang, Y.; Qian, J.; Hassan, M.; Zhang, X.; Zhang, T.; Yang, C.; Zhou, X.; Jia, F. Density peak clustering algorithms: A review on the decade 2014–2023. *Expert Syst. Appl.* **2024**, *238*, 121860. [CrossRef]
24. Parsons, L.; Haque, E.; Liu, H. Subspace clustering for high dimensional data: A review. *SIGKDD Explor.* **2004**, *6*, 90–105. [CrossRef]
25. Ezugwu, A.E.; Shukla, A.K.; Agbaje, M.B.; Oyelade, O.N.; José-García, A.; Agushaka, J.O. Automatic clustering algorithms: A systematic review and bibliometric analysis of relevant literature. *Neural Comput. Appl.* **2020**, *33*, 6247–6306. [CrossRef]
26. Ezugwu, A.E. Nature-inspired metaheuristic techniques for automatic clustering: A survey and performance study. *SN Appl. Sci.* **2020**, *2*, 273. [CrossRef]
27. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
28. Gutiérrez-Salcedo, M.; Martínez, M.Á.; Moral-Munoz, J.A.; Herrera-Viedma, E.; Cobo, M.J. Some bibliometric procedures for analyzing and evaluating research fields. *Appl. Intell.* **2017**, *48*, 1275–1287. [CrossRef]
29. Donthu, N.; Kumar, S.; Mukherjee, D.; Pandey, N.; Lim, W.M. How to conduct a bibliometric analysis: An overview and guidelines. *J. Bus. Res.* **2021**, *133*, 285–296. [CrossRef]
30. Cobo, M.J.; López-Herrera, A.G.; Herrera-Viedma, E.; Herrera, F. science mapping software tools: Review, analysis, and cooperative study among tools. *J. Am. Soc. Inf. Sci. Technol.* **2011**, *62*, 1382–1402. [CrossRef]
31. Aria, M.; Cuccurullo, C. Bibliometrix: An R-tool for comprehensive science mapping analysis. *J. Informetr.* **2017**, *11*, 959–975. [CrossRef]
32. Pranckutė, R. Web of Science (WoS) and Scopus: The titans of bibliographic information in today's academic world. *Publications* **2021**, *9*, 12. [CrossRef]
33. Shiau, W.-L.; Dwivedi, Y.K.; Yang, H.S. Co-citation and cluster analyses of extant literature on social networks. *Int. J. Inf. Manag.* **2017**, *37*, 390–399. [CrossRef]
34. Perianes-Rodriguez, A.; Waltman, L.; van Eck, N.J. Constructing bibliometric networks: A comparison between full and fractional counting. *J. Informetr.* **2016**, *10*, 1178–1195. [CrossRef]
35. van Eck, N.J.; Waltman, L. Citation-based clustering of publications using CitNetExplorer and VOSviewer. *Scientometrics* **2017**, *111*, 1053–1070. [CrossRef]
36. Orduña-Malea, E.; Costas, R. Link-based approach to study scientific software usage: The case of VOSviewer. *Scientometrics* **2021**, *126*, 8153–8186. [CrossRef]
37. Jiang, F.; Liu, G.Z.; Du, J.W.; Sui, Y.F. Initialization of k-modes clustering using outlier detection techniques. *Inf. Sci.* **2016**, *332*, 167–183. [CrossRef]
38. Li, M.; Deng, S.B.; Wang, L.; Feng, S.Z.; Fan, J.P. Hierarchical clustering algorithm for categorical data using a probabilistic rough set model. *Knowl. -Based Syst.* **2014**, *65*, 60–71. [CrossRef]
39. Bai, L.; Liang, J.Y. The k-modes type clustering plus between-cluster information for categorical data. *Neurocomputing* **2014**, *133*, 111–121. [CrossRef]
40. Cao, F.Y.; Huang, J.Z.X.; Liang, J.Y.; Zhao, X.W.; Meng, Y.F.; Feng, K.; Qian, Y.H. An algorithm for clustering categorical data with set-valued features. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 4593–4606. [CrossRef] [PubMed]
41. Qin, H.W.; Ma, X.Q.; Herawan, T.; Zain, J.M. MGR: An information theory based hierarchical divisive clustering algorithm for categorical data. *Knowl. -Based Syst.* **2014**, *67*, 401–411. [CrossRef]
42. Yanto, I.T.R.; Ismail, M.A.; Herawan, T. A modified fuzzy k-partition based on indiscernibility relation for categorical data clustering. *Eng. Appl. Artif. Intell.* **2016**, *53*, 41–52. [CrossRef]
43. McNicholas, P.D. Model-based clustering. *J. Classif.* **2016**, *33*, 331–373. [CrossRef]
44. Goodman, L.A. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **1974**, *61*, 215–231. [CrossRef]
45. Weller, B.E.; Bowen, N.K.; Faubert, S.J. Latent class analysis: A guide to best practice. *J. Black Psychol.* **2020**, *46*, 287–311. [CrossRef]
46. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B (Methodol.)* **2018**, *39*, 1–22. [CrossRef]
47. Wei, W.; Liang, J.Y.; Guo, X.Y.; Song, P.; Sun, Y.J. Hierarchical division clustering framework for categorical data. *Neurocomputing* **2019**, *341*, 118–134. [CrossRef]
48. Sulc, Z.; Rezanková, H. Comparison of similarity measures for categorical data in hierarchical clustering. *J. Classif.* **2019**, *36*, 58–72. [CrossRef]
49. Xu, S.L.; Liu, S.L.; Zhou, J.; Feng, L. Fuzzy rough clustering for categorical data. *Int. J. Mach. Learn. Cybern.* **2019**, *10*, 3213–3223. [CrossRef]

50. Saha, I.; Sarkar, J.P.; Maulik, U. Integrated rough fuzzy clustering for categorical data analysis. *Fuzzy Sets Syst.* **2019**, *361*, 1–32. [CrossRef]
51. Peng, L.W.; Liu, Y.G. Attribute weights-based clustering centres algorithm for initialising k-modes clustering. *Clust. Comput. -J. Netw. Softw. Tools Appl.* **2019**, *22*, S6171–S6179. [CrossRef]
52. Ye, Y.Q.; Jiang, J.; Ge, B.F.; Yang, K.W.; Stanley, H.E. Heterogeneous graph based similarity measure for categorical data unsupervised learning. *IEEE Access* **2019**, *7*, 112662–112680. [CrossRef]
53. Nguyen, T.P.Q.; Kuo, R.J. Automatic fuzzy clustering using non-dominated sorting particle swarm optimization algorithm for categorical data. *IEEE Access* **2019**, *7*, 99721–99734. [CrossRef]
54. Nguyen, T.P.Q.; Kuo, R.J. Partition-and-merge based fuzzy genetic clustering algorithm for categorical data. *Appl. Soft Comput.* **2019**, *75*, 254–264. [CrossRef]
55. Kuo, R.J.; Nguyen, T.P.Q. Genetic intuitionistic weighted fuzzy k-modes algorithm for categorical data. *Neurocomputing* **2019**, *330*, 116–126. [CrossRef]
56. Xiao, Y.Y.; Huang, C.H.; Huang, J.Y.; Kaku, I.; Xu, Y.C. Optimal mathematical programming and variable neighborhood search for k-modes categorical data clustering. *Pattern Recognit.* **2019**, *90*, 183–195. [CrossRef]
57. Gao, X.N.; Wu, S. CUBOS: An internal cluster validity index for categorical data. *Teh. Vjesn. -Tech. Gaz.* **2019**, *26*, 486–494. [CrossRef]
58. Jia, H.; Cheung, Y.M.; Liu, J.M. A new distance metric for unsupervised learning of categorical data. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *27*, 1065–1079. [CrossRef] [PubMed]
59. Qian, Y.H.; Li, F.J.; Liang, J.Y.; Liu, B.; Dang, C.Y. Space structure and clustering of categorical data. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *27*, 2047–2059. [CrossRef] [PubMed]
60. Yang, C.L.; Kuo, R.J.; Chien, C.H.; Quyen, N.T.P. Non-dominated sorting genetic algorithm using fuzzy membership chromosome for categorical data clustering. *Appl. Soft Comput.* **2015**, *30*, 113–122. [CrossRef]
61. Chen, L.F.; Wang, S.R.; Wang, K.J.; Zhu, J.P. Soft subspace clustering of categorical data with probabilistic distance. *Pattern Recognit.* **2016**, *51*, 322–332. [CrossRef]
62. Park, I.K.; Choi, G.S. Rough set approach for clustering categorical data using information-theoretic dependency measure. *Inf. Syst.* **2015**, *48*, 289–295. [CrossRef]
63. Zhu, S.W.; Xu, L.H. Many-objective fuzzy centroids clustering algorithm for categorical data. *Expert. Syst. Appl.* **2018**, *96*, 230–248. [CrossRef]
64. Ben Salem, S.; Naouali, S.; Chtourou, Z. A fast and effective partitional clustering algorithm for large categorical datasets using a k-means based approach. *Comput. Electr. Eng.* **2018**, *68*, 463–483. [CrossRef]
65. Bai, L.; Liang, J.Y. Cluster validity functions for categorical data: A solution-space perspective. *Data Min. Knowl. Discov.* **2015**, *29*, 1560–1597. [CrossRef]
66. Bai, L.; Liang, J.Y. A categorical data clustering framework on graph representation. *Pattern Recognit.* **2022**, *128*, 108694. [CrossRef]
67. Cao, F.Y.; Huang, J.Z.X.; Liang, J.Y. A fuzzy SV-k-modes algorithm for clustering categorical data with set-valued attributes. *Appl. Math. Comput.* **2017**, *295*, 1–15. [CrossRef]
68. Cao, F.Y.; Yu, L.Q.; Huang, J.Z.X.; Liang, J.Y. K-mw-modes: An algorithm for clustering categorical matrix-object data. *Appl. Soft Comput.* **2017**, *57*, 605–614. [CrossRef]
69. Kuo, R.J.; Zheng, Y.R.; Nguyen, T.P.Q. Metaheuristic-based possibilistic fuzzy k-modes algorithms for categorical data clustering. *Inf. Sci.* **2021**, *557*, 1–15. [CrossRef]
70. Ben Salem, S.; Naouali, S.; Chtourou, Z. The DRk-M for clustering categorical datasets with uncertainty. *IEEE Intell. Syst.* **2021**, *36*, 113–121. [CrossRef]
71. Ben Salem, S.; Naouali, S.; Chtourou, Z. A rough set based algorithm for updating the modes in categorical clustering. *Int. J. Mach. Learn. Cybern.* **2021**, *12*, 2069–2090. [CrossRef] [PubMed]
72. Naouali, S.; Ben Salem, S.; Chtourou, Z. Uncertainty mode selection in categorical clustering using the rough set theory. *Expert. Syst. Appl.* **2020**, *158*, 113555. [CrossRef]
73. Zhang, Y.Q.; Cheung, Y.M. A new distance metric exploiting heterogeneous interattribute relationship for ordinal-and-nominal-attribute data clustering. *IEEE Trans. Cybern.* **2022**, *52*, 758–771. [CrossRef] [PubMed]
74. Zhang, Y.Q.; Cheung, Y.M. Learnable weighting of intra-attribute distances for categorical data clustering with nominal and ordinal attributes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 3560–3576. [CrossRef]
75. Zhang, Y.Q.; Cheung, Y.M.; Tan, K.C. A unified entropy-based distance metric for ordinal-and-nominal-attribute data clustering. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 39–52. [CrossRef]
76. Chen, H.; Xu, K.P.; Chen, L.F.; Jiang, Q.S. Self-expressive kernel subspace clustering algorithm for categorical data with embedded feature selection. *Mathematics* **2021**, *9*, 1680. [CrossRef]
77. Chen, L.F. A probabilistic framework for optimizing projected clusters with categorical attributes. *Sci. China-Inf. Sci.* **2015**, *58*, 072104:1–072104:15. [CrossRef]
78. Yuan, F.; Yang, Y.L.; Yuan, T.T. A dissimilarity measure for mixed nominal and ordinal attribute data in k-modes algorithm. *Appl. Intell.* **2020**, *50*, 1498–1509. [CrossRef]
79. Oskouei, A.G.; Balafar, M.A.; Motamed, C. FKMAWCW: Categorical fuzzy k-modes clustering with automated attribute-weight and cluster-weight learning. *Chaos Solitons Fractals* **2021**, *153*, 111494. [CrossRef]

80. Saha, A.; Das, S. Categorical fuzzy k-modes clustering with automated feature weight learning. *Neurocomputing* **2015**, *166*, 422–435. [CrossRef]
81. Heloulou, I.; Radjef, M.S.; Kechadi, M.T. A multi-act sequential game-based multi-objective clustering approach for categorical data. *Neurocomputing* **2017**, *267*, 320–332. [CrossRef]
82. Dorman, K.S.; Maitra, R. An efficient k-modes algorithm for clustering categorical datasets. *Stat. Anal. Data Min.* **2022**, *15*, 83–97. [CrossRef]
83. Rios, E.J.R.; Medina-Pérez, M.A.; Lazo-Cortés, M.S.; Monroy, R. Learning-based dissimilarity for clustering categorical data. *Appl. Sci. -Basel* **2021**, *11*, 3509. [CrossRef]
84. Uddin, J.; Ghazali, R.; Deris, M.M.; Iqbal, U.; Shoukat, I.A. A novel rough value set categorical clustering technique for supplier base management. *Computing* **2021**, *103*, 2061–2091. [CrossRef]
85. Suri, N.; Murty, M.N.; Athithan, G. Detecting outliers in categorical data through rough clustering. *Nat. Comput.* **2016**, *15*, 385–394. [CrossRef]
86. Kar, A.K.; Mishra, A.C.; Mohanty, S.K. An efficient entropy based dissimilarity measure to cluster categorical data. *Eng. Appl. Artif. Intell.* **2023**, *119*, 105795. [CrossRef]
87. Chen, B.G.; Yin, H.T. Learning category distance metric for data clustering. *Neurocomputing* **2018**, *306*, 160–170. [CrossRef]
88. Jian, S.L.; Cao, L.B.; Lu, K.; Gao, H. Unsupervised coupled metric similarity for Non-IID categorical data. *IEEE Trans. Knowl. Data Eng.* **2018**, *30*, 1810–1823. [CrossRef]
89. Zhang, C.B.; Chen, L.; Zhao, Y.P.; Wang, Y.X.; Chen, C.L.P. Graph enhanced fuzzy clustering for categorical data using a bayesian dissimilarity measure. *IEEE Trans. Fuzzy Syst.* **2023**, *31*, 810–824. [CrossRef]
90. Narasimhan, M.; Balasubramanian, B.; Kumar, S.D.; Patil, N. EGA-FMC: Enhanced genetic algorithm-based fuzzy k-modes clustering for categorical data. *Int. J. Bio-Inspired Comput.* **2018**, *11*, 219–228. [CrossRef]
91. Zheng, Q.B.; Diao, X.C.; Cao, J.J.; Liu, Y.; Li, H.M.; Yao, J.N.; Chang, C.; Lv, G.J. From whole to part: Reference-based representation for clustering categorical data. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 927–937. [CrossRef]
92. Faouzi, T.; Firinguetti-Limone, L.; Avilez-Bozo, J.M.; Carvajal-Schiaffino, R. The α-Groups under condorcet clustering. *Mathematics* **2022**, *10*, 718. [CrossRef]
93. Jiang, Z.N.; Liu, X.Y.; Zang, W.K. A kernel-based intuitionistic weight fuzzy k-modes algorithm using coupled chained P system combines DNA genetic rules for categorical data. *Neurocomputing* **2023**, *528*, 84–96. [CrossRef]
94. Amiri, S.; Clarke, B.S.; Clarke, J.L. Clustering categorical data via ensembling dissimilarity matrices. *J. Comput. Graph. Stat.* **2018**, *27*, 195–208. [CrossRef]
95. Kim, K. A weighted k-modes clustering using new weighting method based on within-cluster and between-cluster impurity measures. *J. Intell. Fuzzy Syst.* **2017**, *32*, 979–990. [CrossRef]
96. Sun, H.J.; Chen, R.B.; Qin, Y.; Wang, S.R. Holo-entropy based categorical data hierarchical clustering. *Informatica* **2017**, *28*, 303–328. [CrossRef]
97. Mau, T.N.; Inoguchi, Y.; Huynh, V.N. A novel cluster prediction approach based on locality-sensitive hashing for fuzzy clustering of categorical data. *IEEE Access* **2022**, *10*, 34196–34206. [CrossRef]
98. Dinh, D.T.; Huynh, V.N. k-PbC: An improved cluster center initialization for categorical data clustering. *Appl. Intell.* **2020**, *50*, 2610–2632. [CrossRef]
99. Parmar, D.; Wu, T.; Blackhurst, J. MMR: An algorithm for clustering categorical data using rough set theory. *Data Knowl. Eng.* **2007**, *63*, 879–893. [CrossRef]
100. He, Z.Y.; Xu, X.F.; Deng, S.C. K-ANMI: A mutual information based clustering algorithm for categorical data. *Inf. Fusion.* **2008**, *9*, 223–233. [CrossRef]
101. Deng, S.C.; He, Z.Y.; Xu, X.F. G-ANMI: A mutual information based genetic clustering algorithm for categorical data. *Knowl. -Based Syst.* **2010**, *23*, 144–149. [CrossRef]
102. Barbará, D.; Li, Y.; Couto, J. COOLCAT: An entropy-based algorithm for categorical clustering. In Proceedings of the Eleventh International Conference on Information and Knowledge Management, McLean, VA, USA, 4–9 November 2002; pp. 582–589.
103. Herawan, T.; Deris, M.M.; Abawajy, J.H. A rough set approach for selecting clustering attribute. *Knowl. -Based Syst.* **2010**, *23*, 220–231. [CrossRef]
104. Mazlack, L.; He, A.; Zhu, Y.; Coppock, S. A rough set approach in choosing partitioning attributes. In Proceedings of the ISCA 13th International Conference (CAINE-2000), Honolulu, HI, USA, 1–3 November 2000; pp. 1–6.
105. Andritsos, P.; Tsaparas, P.; Miller, R.J.; Sevcik, K.C. Limbo: A scalable algorithm to cluster categorical data. In Proceedings of the International Conference on Extending Database Technology, Berlin/Heidelberg, Germany, 7–10 December 2003; pp. 123–146.
106. Altameem, A.; Poonia, R.C.; Kumar, A.; Raja, L.; Saudagar, A.K.J. P-ROCK: A sustainable clustering algorithm for large categorical datasets. *Intell. Autom. Soft Comput.* **2023**, *35*, 553–566. [CrossRef]
107. Guha, S.; Rastogi, R.; Shim, K. ROCK: A robust clustering algorithm for categorical attributes. *Inf. Syst.* **2000**, *25*, 345–366. [CrossRef]
108. Wu, S.; Wang, S. Information-theoretic outlier detection for large-scale categorical data. *IEEE Trans. Knowl. Data Eng.* **2013**, *25*, 589–602. [CrossRef]
109. Dutta, M.; Mahanta, A.K.; Pujari, A.K. QROCK: A quick version of the ROCK algorithm for clustering of categorical data. *Pattern Recognit. Lett.* **2005**, *26*, 2364–2373. [CrossRef]

110. Saruladha, K.; Likhitha, P. Modified rock (MROCK) algorithm for clustering categorical data. *Adv. Nat. Appl. Sci.* **2015**, *9*, 518–525.
111. Ben Hariz, S.; Elouedi, Z. New dynamic clustering approaches within belief function framework. *Intell. Data Anal.* **2014**, *18*, 409–428. [CrossRef]
112. Smets, P. The transferable belief model and other interpretations of Dempster-Shafer's model. In Proceedings of the Conference on Uncertainty in Artificial Intelligence, Cambridge, MA, USA, 27–29 July 1990.
113. Ben Hariz, S.; Elouedi, Z.; Mellouli, K. *Clustering Approach Using Belief Function Theory*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 162–171.
114. Cao, F.; Liang, J.; Li, D.; Zhao, X. A weighting k-modes algorithm for subspace clustering of categorical data. *Neurocomputing* **2013**, *108*, 23–30. [CrossRef]
115. Cao, F.; Liang, J.; Li, D.; Bai, L.; Dang, C. A dissimilarity measure for the k-modes clustering algorithm. *Knowl. -Based Syst.* **2012**, *26*, 120–127. [CrossRef]
116. Chi-Hyon, O.; Honda, K.; Ichihashi, H. Fuzzy clustering for categorical multivariate data. In Proceedings of the Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No. 01TH8569), Vancouver, BC, Canada, 25–28 July 2001; Volume 4, pp. 2154–2159.
117. Heloulou, I.; Radjef, M.S.; Kechadi, M.T. *Clustering Based on Sequential Multi-Objective Games*; Springer International Publishing: Munich, Germany, 2014; pp. 369–381.
118. Kaufman, L.; Rousseeuw, P. *Finding Groups in Data: An Introduction to Cluster Analysis*; John Wiley & Sons: New York, NY, USA, 1990.
119. Zhang, M.-L.; Zhou, Z.-H. Multi-instance clustering with applications to multi-instance prediction. *Appl. Intell.* **2009**, *31*, 47–68. [CrossRef]
120. Giannotti, F.; Gozzi, C.; Manco, G. *Clustering Transactional Data*; Springer: Berlin/Heidelberg, Germany, 2002; pp. 175–187.
121. Khan, S.S.; Ahmad, A. Cluster center initialization algorithm for k-modes clustering. *Expert. Syst. Appl.* **2013**, *40*, 7444–7456. [CrossRef]
122. Wu, S.; Jiang, Q.; Huang, J.Z. A new initialization method for clustering categorical data. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Nanjing, China, 22–25 May 2007.
123. Arthur, D.; Vassilvitskii, S. K-means++: The advantages of careful seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, LA, USA, 7–9 January 2007; pp. 1027–1035.
124. Bahmani, B.; Moseley, B.; Vattani, A.; Kumar, R.; Vassilvitskii, S. Scalable k-means++. *Proc. VLDB Endow.* **2012**, *5*, 622–633. [CrossRef]
125. Fuyuan, C.; Jiye, L.; Liang, B. A new initialization method for categorical data clustering. *Expert. Syst. Appl.* **2009**, *36*, 10223–10228. [CrossRef]
126. San, O.M.; Huynh, V.-N.; Nakamori, Y. An alternative extension of the k-means algorithm for clustering categorical data. *Int. J. Appl. Math. Comput. Sci.* **2004**, *14*, 241–247.
127. Nguyen, T.-H.T.; Huynh, V.-N. A k-means-like algorithm for clustering categorical data using an information theoretic-based dissimilarity measure. In Proceedings of the International Symposium on Foundations of Information and Knowledge Systems, Linz, Austria, 7–11 March 2016.
128. Nguyen, T.-H.T.; Dinh, D.-T.; Sriboonchitta, S.; Huynh, V.-N. A method for k-means-like clustering of categorical data. *J. Ambient. Intell. Humaniz. Comput.* **2019**, *14*, 15011–15021. [CrossRef]
129. Nguyen, H.H. Clustering categorical data using community detection techniques. *Comput. Intell. Neurosci.* **2017**, *2017*, 8986360. [CrossRef] [PubMed]
130. Chan, E.Y.; Ching, W.-K.; Ng, M.K.P.; Huang, J.Z. An optimization algorithm for clustering using weighted dissimilarity measures. *Pattern Recognit.* **2004**, *37*, 943–952. [CrossRef]
131. Bai, L.; Liang, J.; Dang, C.; Cao, F. A novel attribute weighting algorithm for clustering high-dimensional categorical data. *Pattern Recognit.* **2011**, *44*, 2843–2861. [CrossRef]
132. Ng, A.Y.; Jordan, M.I.; Weiss, Y. On spectral clustering: Analysis and an algorithm. In Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, Vancouver, BC, Canada, 3–8 December 2001; pp. 849–856.
133. Lee, D.D.; Seung, H.S. Algorithms for non-negative matrix factorization. In Proceedings of the 13th International Conference on Neural Information Processing Systems, Denver, CO, USA, 28–30 November 2000; pp. 535–541.
134. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [CrossRef]
135. Ralambondrainy, H. A conceptual version of the k-means algorithm. *Pattern Recognit. Lett.* **1995**, *16*, 1147–1157. [CrossRef]
136. Iam-On, N.; Boongeon, T.; Garrett, S.; Price, C. A link-based cluster ensemble approach for categorical data clustering. *IEEE Trans. Knowl. Data Eng.* **2012**, *24*, 413–425. [CrossRef]
137. Jian, S.; Cao, L.; Pang, G.; Lu, K.; Gao, H. Embedding-based representation of categorical data by hierarchical value coupling learning. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017; pp. 1937–1943.
138. Marcotorchino, F.; Michaud, P. Agregation de similarites en classification automatique. *Rev. De Stat. Appliquée* **1982**, *30*, 21–44.

139. Hariz, S.B.; Elouedi, Z. IK-BKM: An incremental clustering approach based on intra-cluster distance. In Proceedings of the ACS/IEEE International Conference on Computer Systems and Applications—AICCSA 2010, Washington, DC, USA, 16–19 May 2010; pp. 1–8.
140. Ben Hariz, S.; Elouedi, Z. *DK-BKM: Decremental k Belief k-Modes Method*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 84–97.
141. Hartigan, J.A.; Wong, M.A. A k-means clustering algorithm. *J. R. Stat. Society. Ser. C (Appl. Stat.)* **1979**, *28*, 100–108. [CrossRef]
142. Grahne, G.; Zhu, J. High performance mining of maximal frequent itemsets. In Proceedings of the 6th International Workshop on High Performance Data Mining, San Francisco, CA, USA, 1–3 May 2003; p. 34.
143. Ng, M.K.; Li, M.J.; Huang, J.Z.; He, Z. On the impact of dissimilarity measure in k-modes clustering algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 503–507. [CrossRef]
144. Ben Salem, S.; Naouali, S.; Sallami, M. Clustering categorical data using the k-means algorithm and the attribute's relative frequency. *World Acad. Sci. Eng. Technol. Int. J. Comput. Electr. Autom. Control Inf. Eng.* **2017**, *11*, 708–713.
145. Semeh Ben, S.; Sami, N.; Moetez, S. A computational cost-effective clustering algorithm in multidimensional space using the manhattan metric: Application to the global terrorism database. *World Acad. Sci. Eng. Technol. Int. J. Comput. Electr. Autom. Control Inf. Eng.* **2017**, *2017*, 14.
146. Gan, G.; Wu, J.; Yang, Z. A genetic fuzzy k-Modes algorithm for clustering categorical data. *Expert. Syst. Appl.* **2009**, *36*, 1615–1620. [CrossRef]
147. Mukhopadhyay, A.; Maulik, U.; Bandyopadhyay, S. Multiobjective genetic algorithm-based fuzzy clustering of categorical attributes. *IEEE Trans. Evol. Comput.* **2009**, *13*, 991–1005. [CrossRef]
148. Maciel, D.B.M.; Amaral, G.J.A.; de Souza, R.; Pimentel, B.A. Multivariate fuzzy k-modes algorithm. *Pattern Anal. Appl.* **2017**, *20*, 59–71. [CrossRef]
149. Trigo, M. Using Fuzzy k-Modes to Analyze Patterns of System Calls for Intrusion Detection. Master's Thesis, California State University, Los Angeles, CA, USA, 2005.
150. Kim, D.-W.; Lee, K.H.; Lee, D. Fuzzy clustering of categorical data using fuzzy centroids. *Pattern Recognit. Lett.* **2004**, *25*, 1263–1271. [CrossRef]
151. Cesario, E.; Manco, G.; Ortale, R. Top-down parameter-free clustering of high-dimensional categorical data. *IEEE Trans. Knowl. Data Eng.* **2007**, *19*, 1607–1624. [CrossRef]
152. Tengke, X.; Shengrui, W.; André, M.; Ernest, M. DHCC: Divisive hierarchical clustering of categorical data. *Data Min. Knowl. Discov.* **2012**, *24*, 103–135. [CrossRef]
153. Bouguessa, M. Clustering categorical data in projected spaces. *Data Min. Knowl. Discov.* **2015**, *29*, 3–38. [CrossRef]
154. Potdar, K.; Pardawala, T.; Pai, C. A comparative study of categorical variable encoding techniques for neural network classifiers. *Int. J. Comput. Appl.* **2017**, *175*, 7–9. [CrossRef]
155. Lucasius, C.B.; Dane, A.D.; Kateman, G. On k-medoid clustering of large data sets with the aid of a genetic algorithm: Background, feasiblity and comparison. *Anal. Chim. Acta* **1993**, *282*, 647–669. [CrossRef]
156. Toan Nguyen, M.; Van-Nam, H. Kernel-based k-representatives algorithm for fuzzy clustering of categorical data. In Proceedings of the 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Luxembourg, 11–14 July 2021. [CrossRef]
157. Mau, T.N.; Huynh, V.N. An LSH-based k-representatives clustering method for large categorical data. *Neurocomputing* **2021**, *463*, 29–44. [CrossRef]
158. Tao, X.; Wang, R.; Chang, R.; Li, C. Density-sensitive fuzzy kernel maximum entropy clustering algorithm. *Knowl. -Based Syst.* **2019**, *166*, 42–57. [CrossRef]
159. Teng, Y.; Qi, S.; Han, F.; Xu, L.; Yao, Y.; Qian, W. Two graph-regularized fuzzy subspace clustering methods. *Appl. Soft Comput.* **2021**, *100*, 106981. [CrossRef]
160. Pal, N.R.; Pal, K.; Keller, J.M.; Bezdek, J.C. A possibilistic fuzzy c-means clustering algorithm. *IEEE Trans. Fuzzy Syst.* **2005**, *13*, 517–530. [CrossRef]
161. Chaudhuri, A. Intuitionistic fuzzy possibilistic c means clustering algorithms. *Adv. Fuzzy Syst.* **2015**, *2015*, 238237. [CrossRef]
162. Xu, D.; Xu, Z.; Liu, S.; Zhao, H. A spectral clustering algorithm based on intuitionistic fuzzy information. *Knowl. -Based Syst.* **2013**, *53*, 20–26. [CrossRef]
163. Xu, Z.; Chen, J.; Wu, J. Clustering algorithm for intuitionistic fuzzy sets. *Inf. Sci.* **2008**, *178*, 3775–3790. [CrossRef]
164. Zeshui, X. Intuitionistic fuzzy hierarchical clustering algorithms. *J. Syst. Eng. Electron.* **2009**, *20*, 90–97.
165. Păun, G. Computing with membranes. *J. Comput. Syst. Sci.* **2000**, *61*, 108–143. [CrossRef]
166. Zang, W.; Sun, M.; Jiang, Z. A DNA genetic algorithm inspired by biological membrane structure. *J. Comput. Theor. Nanosci.* **2016**, *13*, 3763–3772. [CrossRef]
167. Ammar, A.; Elouedi, Z.; Lingras, P. Semantically segmented clustering based on possibilistic and rough set theories. *Int. J. Intell. Syst.* **2015**, *30*, 676–706. [CrossRef]
168. Tripathy, B.K.; Ghosh, A. SDR: An algorithm for clustering categorical data using rough set theory. In Proceedings of the 2011 IEEE Recent Advances in Intelligent Computational Systems, Trivandrum, India, 22–24 September 2011; pp. 867–872.
169. Tripathy, B.K.; Adhir, G. SSDR: An algorithm for clustering categorical data using rough set theory. *Adv. Appl. Sci. Res.* **2011**, *2*, 314–326.
170. Yang, M.-S.; Chiang, Y.-H.; Chen, C.-C.; Lai, C.-Y. A fuzzy k-partitions model for categorical data and its comparison to the GoM model. *Fuzzy Sets Syst.* **2008**, *159*, 390–405. [CrossRef]

171. Zengyou, H.; Xiaofei, X.; Shengchun, D. A cluster ensemble method for clustering categorical data. *Inf. Fusion.* **2005**, *6*, 143–151. [CrossRef]
172. Ng, M.K.; Wong, J.C. Clustering categorical data sets using tabu search techniques. *Pattern Recognit.* **2002**, *35*, 2783–2790. [CrossRef]
173. Jain, A.K.; Dubes, R.C. *Algorithms for Clustering Data*; Prentice-Hall, Inc.: Saddle River, NJ, USA, 1988.
174. Saha, I.; Sarkar, J.P.; Maulik, U. Ensemble based rough fuzzy clustering for categorical data. *Knowl. -Based Syst.* **2015**, *77*, 114–127. [CrossRef]
175. Peters, G.; Lampart, M.; Weber, R. Evolutionary rough k-medoid clustering. In *Transactions on Rough Sets VIII*; Peters, J.F., Skowron, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2008; pp. 289–306.
176. Huang, J.Z.; Ng, M.K.; Hongqiang, R.; Zichen, L. Automated variable weighting in k-means type clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 657–668. [CrossRef]
177. Qin, H.; Ma, X.; Zain, J.M.; Herawan, T. A novel soft set approach in selecting clustering attribute. *Knowl. -Based Syst.* **2012**, *36*, 139–145. [CrossRef]
178. Bai, L.; Liang, J.; Dang, C.; Cao, F. A novel fuzzy clustering algorithm with between-cluster information for categorical data. *Fuzzy Sets Syst.* **2013**, *215*, 55–73. [CrossRef]
179. Hassanein, W.A.; Elmelegy, A.A. An algorithm for selecting clustering attribute using significance of attributes. *Int. J. Database Theory Appl.* **2013**, *6*, 53–66. [CrossRef]
180. Ammar, A.; Elouedi, Z.; Lingras, P. The k-modes method using possibility and rough set theories. In Proceedings of the 2013 Joint IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), Edmonton, AB, Canada, 24–28 June 2013; pp. 1297–1302.
181. Lee, J.; Lee, Y.J. An effective dissimilarity measure for clustering of high-dimensional categorical data. *Knowl. Inf. Syst.* **2014**, *38*, 743–757. [CrossRef]
182. Tao, L.; Sheng, M.; Mitsunori, O. Entropy-based criterion in categorical clustering. In Proceedings of the Twenty-First International Conference on Machine Learning, Banff, AB, Canada, 4–8 July 2004. [CrossRef]
183. Liang, B.; Jiye, L.; Chuangyin, D.; Fuyuan, C. The impact of cluster representatives on the convergence of the k-modes type clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1509–1522. [CrossRef]
184. Esposito, F.; Malerba, D.; Tamma, V.; Bock, H.-H. *Classical Resemblance Measures*; Springer: Berlin/Heidelberg, Germany, 2000; pp. 139–152.
185. Ahmad, A.; Dey, L. A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set. *Pattern Recognit. Lett.* **2007**, *28*, 110–118. [CrossRef]
186. Knorr, E.M.; Ng, R.T. Algorithms for mining distance-based outliers in large datasets. In Proceedings of the Very Large Data Bases Conference, New York, NY, USA, 24–27 August 1998.
187. Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2–4 August 1996; pp. 226–231.
188. Fraley, C.; Raftery, A.E. Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.* **2002**, *97*, 611–631. [CrossRef]
189. Ahmad, A.; Dey, L. A k-mean clustering algorithm for mixed numeric and categorical data. *Data Knowl. Eng.* **2007**, *63*, 503–527. [CrossRef]
190. Wang, C.; Cao, L.; Wang, M.; Li, J.; Wei, W.; Ou, Y. Coupled nominal similarity in unsupervised learning. In Proceedings of the 20th ACM international conference on Information and knowledge management, Glasgow, Scotland, 24–28 October 2011; pp. 973–978.
191. Wang, C.; Dong, X.; Zhou, F.; Cao, L.; Chi, C.-H. Coupled Attribute Similarity learning on categorical data. *IEEE Trans. Neural Netw. Learn. Syst.* **2015**, *26*, 781–797. [CrossRef]
192. Boriah, S.; Chandola, V.; Kumar, V. Similarity measures for categorical data: A comparative evaluation. In Proceedings of the 2008 SIAM International Conference on Data Mining (SDM); SIAM: Atlanta, GA, USA, 2008; pp. 243–254.
193. Bock, H.-H.; Diday, E. *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*; Springer Science & Business Media: Munich, Germany, 2000.
194. von Luxburg, U. A tutorial on spectral clustering. *Stat. Comput.* **2007**, *17*, 395–416. [CrossRef]
195. Jones, K.S. A statistical interpretation of term specificity and its application in retrieval. In *Document Retrieval Systems*; Taylor Graham Publishing: Abingdon, UK, 1988; pp. 132–142.
196. David, W.G. A new similarity index based on probability. *Biometrics* **1966**, *1966*, 882–907. [CrossRef]
197. Li, C.; Li, H. A modified short and fukunaga metric based on the attribute independence assumption. *Pattern Recognit. Lett.* **2012**, *33*, 1213–1218. [CrossRef]
198. Eskin, E.; Arnold, A.; Prerau, M.; Portnoy, L.; Stolfo, S. A geometric framework for unsupervised anomaly detection. In *Applications of Data Mining in Computer Security*; Barbará, D., Jajodia, S., Eds.; Springer: Boston, MA, USA, 2002; pp. 77–101.
199. Morlini, I.; Zani, S. A new class of weighted similarity indices using polytomous variables. *J. Classif.* **2012**, *29*, 199–226. [CrossRef]
200. Lin, D. An information-theoretic definition of similarity. In Proceedings of the Fifteenth International Conference on Machine Learning, Wisconson, DC, USA, 24–27 July 1998; pp. 296–304.
201. Sokal, R.R.; Michener, C.D. A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.* **1958**, *38*, 1409–1438.

202. Dino, I.; Ruggero, G.P.; Rosa, M. *Context-Based Distance Learning for Categorical Data Clustering*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 83–94.
203. Dino, I.; Ruggero, G.P.; Rosa, M. From context to distance: Learning dissimilarity for categorical data clustering. *ACM Trans. Knowl. Discov. Data* **2012**, *6*, 1. [CrossRef]
204. Liping, J.; Michael, K.N.; Joshua Zhexue, H. An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Trans. Knowl. Data Eng.* **2007**, *19*, 1026–1041. [CrossRef]
205. Jia, H.; Cheung, Y.-m. Subspace clustering of categorical and numerical data with an unknown number of clusters. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 3308–3325.
206. Jian, S.L.; Pang, G.S.; Cao, L.B.; Lu, K.; Gao, H. CURE: Flexible categorical data representation by hierarchical coupling learning. *IEEE Trans. Knowl. Data Eng.* **2019**, *31*, 853–866. [CrossRef]
207. Zhu, C.; Cao, L.; Yin, J. Unsupervised heterogeneous coupling learning for categorical representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 533–549. [CrossRef]
208. Zhang, Y.; Cheung, Y.-m. An ordinal data clustering algorithm with automated distance learning. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 6869–6876. [CrossRef]
209. Murthy, K.P.N. Ludwig boltzmann, transport equation and the second law. *arXiv* **2006**, arXiv:cond-mat/0601566.
210. Du, M.; Ding, S.; Xue, Y. A novel density peaks clustering algorithm for mixed data. *Pattern Recognit. Lett.* **2017**, *97*, 46–53. [CrossRef]
211. Hamming, R.W. Error detecting and error correcting codes. *Bell Syst. Tech. J.* **1950**, *29*, 147–160. [CrossRef]
212. Gambaryan, P. A mathematical model of taxonomy. *Izvest. Akad. Nauk. Armen. SSR* **1964**, *17*, 47–53.
213. Burnaby, T.P. On a method for character weighting a similarity coefficient, employing the concept of information. *J. Int. Assoc. Math. Geol.* **1970**, *2*, 25–38. [CrossRef]
214. Chatzis, S.P. A fuzzy c-means-type algorithm for clustering of data with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional. *Expert. Syst. Appl.* **2011**, *38*, 8684–8689. [CrossRef]
215. de Amorim, R.C.; Makarenkov, V. Applying subclustering and Lp distance in weighted k-means with distributed centroids. *Neurocomputing* **2016**, *173*, 700–707. [CrossRef]
216. Mahamadou, A.J.D.; Antoine, V.; Nguifo, E.M.; Moreno, S. Categorical fuzzy entropy c-means. In Proceedings of the 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Glasgow, UK, 19–24 July 2020; pp. 1–6.
217. Huang, J.Z.; Ng, M.K. A fuzzy k-modes algorithm for clustering categorical data. *IEEE Trans. Fuzzy Syst.* **1999**, *7*, 446–452. [CrossRef]
218. Hashemzadeh, M.; Oskouei, A.G.; Farajzadeh, N. New fuzzy C-means clustering method based on feature-weight and cluster-weight learning. *Appl. Soft Comput.* **2019**, *78*, 324–345. [CrossRef]
219. Zhi, X.B.; Fan, J.L.; Zhao, F. Robust local feature weighting hard c-means clustering algorithm. *Neurocomputing* **2014**, *134*, 20–29. [CrossRef]
220. He, Z.; Deng, S.; Xu, X. *Improving k-Modes Algorithm Considering Frequencies of Attribute Values in Mode*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 157–162.
221. Huang, J.Z. A fast clustering algorithm to cluster very large categorical data sets in data mining. In Proceedings of the Data Mining and Knowledge Discovery, Tucson, AZ, USA, 11 May 1997.
222. Gluck, M.; Corter, J. Information uncertainty, and the utility of categories. In Proceedings of the Seventh Annual Conference of the Cognitive Science Society, Irvine, CA, USA, 15–17 August 1985; pp. 283–287.
223. Gao, C.; Pedrycz, W.; Miao, D.Q. Rough subspace-based clustering ensemble for categorical data. *Soft Comput.* **2013**, *17*, 1643–1658. [CrossRef]
224. Chang, C.-H.; Ding, Z.-K. *Categorical Data Visualization and Clustering Using Subjective Factors*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 229–238.
225. Michaud, P. Clustering techniques. *Future Gener. Comput. Syst.* **1997**, *13*, 135–147. [CrossRef]