Evolving Systems
https://doi.org/10.1007/s12530-023-09503-2

ORIGINAL PAPER

Concept drift from 1980 to 2020: a comprehensive bibliometric analysis with future research insight

Elif Selen Babüroğlu[1] · **Alptekin Durmuşoğlu**[1] · **Türkay Dereli**[2]

Received: 16 September 2022 / Accepted: 7 April 2023

Abstract
In nonstationary environments, high-dimensional data streams have been generated unceasingly where the underlying distribution of the training and target data may change over time. These drifts are labeled as *concept drift* in the literature. Learning from evolving data streams demands adaptive or evolving approaches to handle concept drifts, which is a brand-new research affair. In this effort, a wide-ranging comparative analysis of concept drift is represented to highlight state-of-the-art approaches, embracing the last four decades, namely from 1980 to 2020. Considering the scope and discipline; the core collection of the Web of Science database is regarded as the basis of this study, and 1,564 publications related to concept drift are retrieved. As a result of the classification and feature analysis of valid literature data, the bibliometric indicators are revealed at the levels of countries/regions, institutions, and authors. The overall analyses, respecting the publications, citations, and cooperation of networks, are unveiled not only the highly authoritative publications but also the most prolific institutions, influential authors, dynamic networks, etc. Furthermore, deep analyses including text mining such as; the burst detection analysis, co-occurrence analysis, timeline view analysis, and bibliographic coupling analysis are conducted to disclose the current challenges and future research directions. This paper contributes as a remarkable reference for invaluable further research of concept drift, which enlightens the emerging/trend topics, and the possible research directions with several graphs, visualized by using the VOS viewer and Cite Space software.

Keywords Concept drift · Data stream · Text mining · Bibliometric · Cite Space · VOS viewer

1 Introduction

In 2020; from all around the world, 306.4 billion emails are sent, 5 million tweets are posted on Twitter and 95 million photos and videos are shared on Instagram, each day. Through these transactions, a massive amount of data has been generated continuously; from the Internet of Things realms (IoT), sensor applications, network monitoring, telecommunications, banking, log records or click-streams in web exploring, social media, digital marketing, highway traffic flow monitoring, smartphones, assistive technologies such as Amazon's Alexa, etc., (Bifet et al. 2010, 2019; Mahdi et al. 2020).

Such data are characterized as unbounded, high-speed, high-dimensional, and swiftly evolving (Abdallah et al. 2016; Khamassi et al. 2018; Li et al. 2020). In conjunction with the rapidly growing digital universe, the widespread and repeated dissemination of streaming data in many critical real-time tasks, such as fraud detection, e-mail spam filtering, shopping records, weather sensors or economical prediction (Nordahl et al. 2022), has made evolving data streams a hot research topic (Ren et al. 2018; Babüroğlu et al. 2021) in the knowledge discovery research area (Anupama and Jena 2019), receiving growing attention in machine learning and data mining communities (Santos et al. 2019). Learning from these evolving data streams is a challenging task because of not only managing memory and time efficiently but also possible changes in the underlying data distribution (Baena-Garcia et al. 2006; Bifet and Gavaldà

✉ Elif Selen Babüroğlu
esbuyuknacar@gantep.edu.tr

Alptekin Durmuşoğlu
durmusoglu@gantep.edu.tr

Türkay Dereli
turkay.dereli@hku.edu.tr

[1] Department of Industrial Engineering, Faculty of Engineering, Gaziantep University, 27310 Gaziantep, Turkey

[2] Hasan Kalyoncu University, Gaziantep, Turkey

Published online: 05 May 2023

2007; Krawczyk and Woźniak 2015), and often leading to an additional constraint stating that each instance of data can be read-only once (Krawczyk et al. 2017; Hidalgo et al. 2019).

For instance, in the e-mail spam filtering problem, it is more pragmatic to assume that, all emails are not available at first but appear in a timely fashion. Every email is treated as a test example formerly and labeled (spam or ham) by an expert, whereas the time to predict a label is also strictly limited, due to the real-time characteristics of data streams (Heusinger et al. 2022). Then it is treated as a training example and used to update the classifier (Hosseini et al. 2013). Since the previous data are no longer attainable and unlabeled data is much larger than that of the labeled data, unsupervised learning should be promoted to discover hidden patterns in streaming data (Elwell and Polikar 2011; Gözüaçık and Can 2021). The phenomenon of underlying data distribution change is known as *concept drift* in the literature (Schlimmer and Granger 1986; Widmer and Kubat 1996; Elwell and Polikar 2009), meaning a statistically significant difference between the joint probability of input and output variables observed in different dataset samples (Giusti et al. 2022), and very pervasive in real-world applications (Dong et al. 2018).

The accuracy of classification decreases as concept drifts arise, which emerges that learners should be adaptive to dynamic changes (Wang et al. 2018b). For instance, users might change their areas of interest or preferences over time in an information system. Hence, learning algorithms employed to form a pattern for the real-world streaming data environment must be able to adapt rapidly and precisely to potential changes (Minku et al. 2010), to lessen the classification error rate. As a result, adaptation to novel distributions is essential to procure the efficiency of the decision-making process. Adaptive learning indicates the adaptation of predictive models online, in response to unforeseen concept drifts. The algorithms of adaptive learning, called state-of-the-art incremental learning, are adapted to the evolution of the data-generating process over time (Gama et al. 2014). An undesirable deterioration in the performance of the algorithms, such as prediction accuracy, might be originated from the failure of adaptation to drifts (Wang and MacHida 2021).

In the non-stationary environment, handling concept drift might affect a vast range of disciplines and domains (Plamen et al. 2010; Lughofer and Angelov 2011; Henzgen et al. 2014; Pratama et al. 2017). Evolving Intelligent Systems (EIS) provides a unique solution for concept drift in streaming data which is a strictly one-pass learning procedure to cope with time-critical applications (Abdullatif et al. 2018; Suárez-Cetrulo et al. 2023), where data streams are generated at a rapid sampling rate. Adaptive and evolving learning algorithms are the most appropriate methods for this challenging issue. The adaptive learning algorithms explicitly handle drift by incorporating forgetting mechanisms and eliminating data from primitive concepts. The aspiration is to update the prior knowledge and adjust the learning models to react to the changes properly. The evolving learning models, which are inherently able to handle concept drifts in streaming data, may be considered an extension of incremental algorithms. These algorithms inevitably follow the movement of data distributions to evolve the model structure and provide a smooth transition from an out-dated version model to an incoming one.

Another way for handling concept drift, drift detection methods, as the principal component of adaptive learning algorithms, are aimed to substitute the base classifier after detecting the drift in the probability distribution of the data improving overall accuracy (Barros and Santos 2018; Souto et al. 2019), with minimum delay. Further, the detection algorithms have to abstain from high false positive and false negative rates as they process input data (Sakthithasan et al. 2013). False-positive rates signify false alarms for concept drift, whereas false-negative means missing the real concept drifts out. False-positive entails keeping resources busier, whereas false-negative causes loss in classification accuracy (Gama et al. 2004, 2014; Bifet and Gavaldà 2007; Huang et al. 2015). In short, the main challenge is to detect and alarm the drift as rapidly as possible, with low false-positive and false-negative rates, distinct from the valid data distribution function.

The related work presented that the papers are concerned to detect, classify and handle concept drift, in the literature. Additionally, several survey papers on concept drift, are printed. The most well-known and referenced review is published by Gama et al. (Gama et al. 2014), which explicitly clarifies the concept drift handling systems in step with adaptive learning. Later, a review paper (Lu et al. 2019), evaluated up-to-date developments on concept drift handling methodologies, while introducing the term of "concept drift understanding" as a new component of learning under concept drift. In 2015, a comprehensive survey has been published by Ditzler et al. This paper (Ditzler et al. 2015) evaluates and compares adaptive and evolving state-of-the-art approaches by highlighting two perspectives; active and passive.

A review paper (Hu et al. 2019) is drawing attention to different characteristics of concept drift and their categorization, likewise another review paper (Iwashita and Papa 2019) concerns with different types of drifts and approaches to handling such changes in the data. Other related surveys (Khamassi et al. 2018; Wares et al. 2019) are also categorized the existing handling approaches of concept drift, by providing comparative analysis on methods. In addition to these review papers, some publications explored handling concept drift in specific learning tasks. For instance, an extensive overview of approaches is provided by a survey

(Gemaque et al. 2020) to tackle concept drift in classification problems in an unsupervised manner, whereas the supervised concept drift detectors are also reviewed by a large-scale comparison (Barros and Santos 2018).

There are other papers related to class-imbalanced data streams (Hoens et al. 2011; Wang et al. 2018b) in the literature, which overviews the methods learning from and adapting to non-stationary environment. A comprehensive survey based on ensemble learning (Krawczyk et al. 2017) focuses on the taxonomy of ensemble algorithms for data stream mining tasks in dynamic environments. In recent past, a literature review paper of drifted stream mining (Agrahari and Singh 2021) is introduced to suggest research trends and categorized the concept drift detectors broadly. Nevertheless, despite the availability of detailed review papers in concept drift (Bayram et al. 2022), none of the them have explored the hot spots and cutting-edge trends of concept drift via in-depth systematic research, to fulfill the need for discovering the current status and forecasting the potential future directions.

This paper purposes a comprehensive bibliometric analysis to uncover the hot topics in the concept drift area based upon the core collection database of "Web of Science" for the last four decades, namely from 1980 to 2020. For visualizing the valid literature, the VOS viewer (van Eck and Waltman 2010), and Cite Space (Synnestvedt et al. 2005) software are employed which are popular scientific knowledge mapping tools. The analysis of bibliometrics is an appliance to evaluate the merits of an explicit field of research and mapping science may view the framework of the area indeed (Cobo et al. 2011; White 2018).

Bibliometrics is a comparatively mature and crucial member of intelligence science (Borgman and Furner 2002; Wang et al. 2018a), and also an effective method that is grounded on quantitative analysis, and the blend of mathematics, statistics, linguistics, and information science, in a distinct area (He et al. 2017; Wang et al. 2021). Bibliometrics has been broadly used in diverse areas because of the capability of discovering the internal structure and the growth direction of a specific research trend or a particular journal. For example, the structure and the evolution of a particular journal; the Information Sciences (Yu et al. 2017; Merigó et al. 2018).

The contributions of the paper can be summarized as; (a) the literature quantitative analysis is introduced to designate the current status of "concept drift researches" based on three different categories; countries or regions, institutions, and authors, regarding the basic bibliometric indicators, such as H-index (Alonso et al. 2009), the sum of publications and citations counted, the number of citations per publication, etc., (b) cooperation networks analysis demonstrates the evolution of concept drift with regards to authors, institutions, countries/regions, and their relations, (c) the hot/trend topics, current challenges, and potential future directions of concept drift area are represented with the aid of analyses such as; timeline view, co-occurrence, burst detection, and bibliographic coupling.

The remainder of this article is structured as follows. The scope of this study is clarified in Sect. 2 and also the method, and data source; Sect. 3 consists of the overall analysis of "concept drift" with four subsections; Sect. 4 presents the cooperation networks of concept drift; Sect. 5 expounds on the in-depth analyses including burst detection analysis, co-occurrence analysis, timeline view analysis and bibliographic coupling analysis and Sect. 6 concludes the paper, providing further discussions on the current challenges and possible future research directions.

2 Scope, method, and data source

Due to the ubiquity in a large spectrum of real-world applications; evolving data streams have blossomed out as a trend research topic. Learning from streaming data, which dynamically derives over time; often in unforeseen ways, is a challenging task.

Numerous scientific papers have been published, to validate, detect, or handle the drift in data. In this study, a bibliometric overview of concept drift during the last four decades, namely from 1980 to 2020, is illustrated. The databases, such as Web of Science, Google Scholar, Scopus, Microsoft Academic, Derwent, PubMed, ADS, arXiv, etc., can be regarded as the data source. In consideration of the scope and the discipline; the core collection of the Web of Science database, which is one of the most widely-used databases in academics (Falagas et al. 2008), is preferred as the basis of this study.

Web of Science (WoS) maintains a huge amount of inclusive data and elaborative information related to publications all over the world, which consists of a broad range of high-quality collections of journals available, covering the indices of; (SCI-E) Science Citation Index Expanded, (SSCI) Social Sciences Citation Index, (A&HCI) Arts and Humanities Citation Index, (ESCI) Emerging Sources Citation Index, and (CPCI) Conference Proceedings Citation Index.

The portal of WoS employs a basic search method, and the keyword "concept drift" is set for the complete retrieval of relevant literature. In between the customized year range, from 1980 to 2020, 1566 publications have been obtained from WoS with the inquiry made on December 13, 2021. The research master files are exported, in both plain text format and tab-delimited format with full records and cited references, to accommodate the entire bibliographic information.

While practicing with this type of complex data which may contain repetitive information such as authors' names, preprocessing is crucial to improve the accuracy of the analysis. Thus, Cite Space is engaged to examine the intricate data retrieved from Web of Science, which is a well-known software that discards duplicated data and demonstrates the characteristics of a related research area. Accordingly, 1564 concept drift-related publications are collected after the data purging.

The records of these 1564 documents are handled as the original research data to generate diverse informative graphs. The other effective bibliometric software tool, VOS viewer (Version 1.6.16), developed by Nees Jan van Eck and Ludo Waltman, is used to exhibit the present standing and growing tendency of a specific topic through graphs.

3 Overall analysis of "concept drift"

At first, the overall analyses for each bibliometric indicator, concerning the countries/regions, institutions, and authors, have been done by considering the number and type of the papers, and the total/average number of citations. The fundamental contribution of the analysis is to explore the hot topics in the concept drift area and future research directions, by introducing the highly authoritative publications, most prolific institutions, influential authors, etc.

3.1 Classification of publications

Following the inquiry of the Web of Science database, the concept drift-related publications are separated into eight particular categories, represented in Fig. 1. The type of proceedings paper engages a relatively greater proportion of all documents with the number of 876.

The publication type article is the second-largest shareholder with 703 records. The reviews take the third row with 26 printed materials. Additionally, there are 3 early access, 10 book chapters, 3 editorial material, 2 corrections, and 1 meeting abstract. In short, the article and proceeding paper types of publications have been the major alternatives for researchers.

The head research directions of the publications attributed to concept drift are illustrated in Fig. 2. Computer science (1385) and engineering (442) are the outstanding research aspects that take the largest piece of the pie. They are chased by particular research directions such as; telecommunications (109), operations research and management science (68), mathematics (47), automation control systems (49), science technology other topics (26), robotics (21), chemistry (16), and business economics (16).

These directions indicated that abundant research harvests are based on the theoretical foundation, like computer science, engineering, telecommunications, operations research, automation control systems, mathematics, science technology, and robotics.

The reason why is an urgent need for handling concept drift in streaming data with adequate drift detection methods to increase the accuracy of the dataset and tests.

3.2 Prolific countries/regions

Feature analysis of valid literature data ascertained that there are 80 different regions/countries. In this manner, a regional classification assists the researcher to explore the geographical pattern rapidly and also to gain insight into the

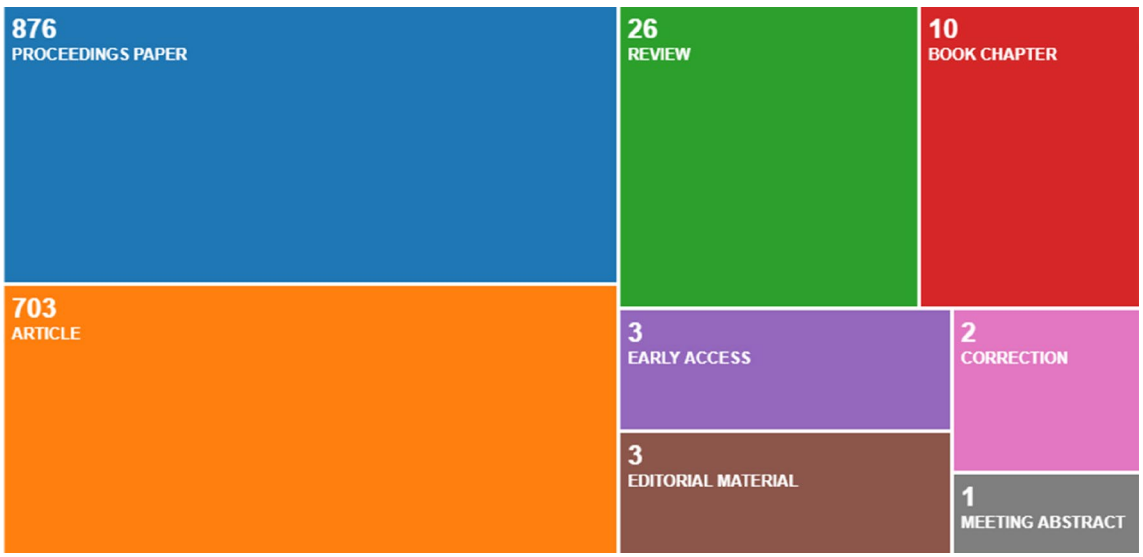

Fig. 1 Sort of publications

Fig. 2 The head 10 research directions of the publications

strength of each country/region in the research field. Regarding the aspects of countries/regions; 1564 documents are evaluated to designate the most productive ones, considering these distinct indicators; (1) the total number of publications (TP), (2) the total number of citations (TC), (3) the average number of citations per publication (TC/TP), 4) the percent of TP accounting for total publications (%TP), 5–6–7) the number of publications that are cited more than or equal to 300/100/50 ($\geq 300/\geq 100/\geq 50$), and 8) H-index, proposed by Hirsch.

The h-index of the researcher accounts for the intersection point of the number of citations and the order of publication (Alonso et al. 2009). The indicators, commonly used in the bibliometric analysis, manifest the current status of publications.

Concerning the sum of publications, the ten most advantageous countries/regions are shown in Table 1, conforming to proportional statistical analysis of indicators. The most prolific countries/regions are aligned with the total number of publications (TP) and the percent of TP accounting for total publications (%TP), as shown in Table 1. Considering TP and %TP; China is the most fruitful country/region with 297 papers.

The USA has the second-highest score with 262 publications, followed by Brazil which has 8.44% of papers with 132 publications. Portugal is ranked as the 10th most productive country/region by having 4.35% of papers. Nonetheless, Portugal has the utmost ratio of TC/TP (52.42), indicating that Portugal has received the highest recognition per publication. In consideration of TC/TP, Spain follows Portugal with a ratio of 36.29 and the third-best score ratio is (36.22), which belongs to England. It's proved that; China, the USA, and Brazil performed well in concept drift-related publications, whereas European countries/regions, such as England, Portugal, or Spain, have been recognized by other countries/regions to a great extent.

Even the total number of citations (TC) provides insights about the characteristics and popularity of the publications,

Table 1 The 10 most prolific countries/regions

Rank	Country/region	TP	TP	TC/TP	%TP	≥ 300	≥ 100	≥ 50	H-index
1	Peoples R China	297	2269	7.640	18.99	0	0	5	24
2	USA	262	5595	21.35	16.75	4	8	12	35
3	Brazil	132	1703	5.325	8.44	0	3	3	21
4	England	105	3803	36.22	6.71	3	6	5	23
5	Poland	103	1706	16.56	6.59	1	2	5	16
6	India	91	248	2.73	5.82	0	0	0	9
7	Australia	89	1370	15.39	5.70	0	4	2	19
8	Germany	89	790	8.876	5.70	0	1	3	14
9	Spain	84	3048	36.29	5.37	2	5	1	20
10	Portugal	68	3564	52.42	4.35	2	1	1	21

some papers were cited more than 300 times; which are called "highly authoritative publications". The detailed information of the 10 head highly authoritative publications is demonstrated in Table 2. There are seven articles, and three proceedings papers listed. The USA is evaluated as the largest share-holder with 4 (four) articles published, among the top 10 highly authoritative publications cited more than 300 times. 2 publications out of these 10 papers (1 proceedings paper and 1 article) come from Portugal, followed by Israel with 2 publications (1 proceedings paper and 1 article). An article, ranked as the second-highest citation/year ratio, comes from Austria; followed by one proceedings paper from Spain.

The number of publications related to each country/region is separated and displayed in Fig. 3. China has Benjamin's portion with a great number of publications (297) regarding the total number of publications (1564), and it is worth considering that the large population might be an upper hand for China. Nevertheless, the USA has a greater proportion of the number of citations as shown in Fig. 4., which represents the citations of publications for the head 10 countries/regions.

Regarding Figs. 3 and 4, it is obvious that England has higher recognition per publication than Brazil, even though Brazil is more productive than England. Spain has a remarkable impact on other publications, with fewer publications than Poland. The cooperation of publications from India seemed a bit restricted for a reason, even though the amount of output is high which ranked in the 6th row.

3.3 Prolific institutions

The allocation of research and development institutions based on concept drift-related research aids the researchers to cooperate, compare and contrast better. Likewise, the prolific countries/regions are investigated according to five distinct indicators, employed to examine 1,240 institutions. The most prolific 10 institutions are listed in Table 3, and the total number of publications of the head 10 institutions is represented in Fig. 5. Universidade do Porto ranks the first in the sense of each indicator such as; TP, TC, TP/TC, TP%, and H-index, with 49 unique publications, which means that the university has a wide range of recognition of concept drift.

The TC/TP ratios are correlated with the number of publications, proving that Universidade do Porto has the highest degree (368.2), pursued by Rowan University (57.35), the University of Waikato (29.37), Politecnico di Milano (25.85), and Wroclaw University of Science and Technology (25.83). Although the Nanyang Technological

Table 2 The details of the top 10 highly authoritative publications

Rank	Title	Author	Year	Citation	Citation/Year	Document Type	Country/Region
1	A Survey on Concept Drift Adaptation	Gama, J; Zliobaite, I; Bifet, A; Pechenizkiy, M; Bouchachia, A	2014	1005	125.63	Article	Portugal
2	Learning in the presence of concept drift and hidden contexts	Widmer, G; Kubat, M	1996	918	35.31	Article	Austria
3	Learning with drift detection	Gama, J; Medas, P; Castillo, G; Rodrigues, P	2004	637	35.39	Proceedings Paper	Portugal
4	Collaborative Filtering with Temporal Dynamics	Koren, Y	2010	474	39.50	Article	Israel
5	Incremental Learning of Concept Drift in Nonstationary Environments	Elwell, Ryan; Polikar, Robi	2011	417	37.91	Article	USA
6	Learning from Time-Changing Data with Adaptive Windowing	Bifet, A; Gavalda, R	2007	384	25.60	Proceedings Paper	Spain
7	Collaborative Filtering with Temporal Dynamics	Koren, Yehuda	2009	364	28.00	Proceedings Paper	Israel
8	Dynamic weighted majority: An ensemble method for drifting concepts	Kolter, J. Zico; Maloof, Marcus A	2007	363	24.20	Article	USA
9	Ensemble learning for data stream analysis: A survey	Krawczyk, B; Minku, LL; Gama, J; Stefanowski, J; Wozniak, M	2017	349	69.80	Article	USA
10	Learning in Nonstationary Environments: A Survey	Ditzler, G; Roveri, M; Alippi, C; Polikar, R	2015	315	45.00	Article	USA

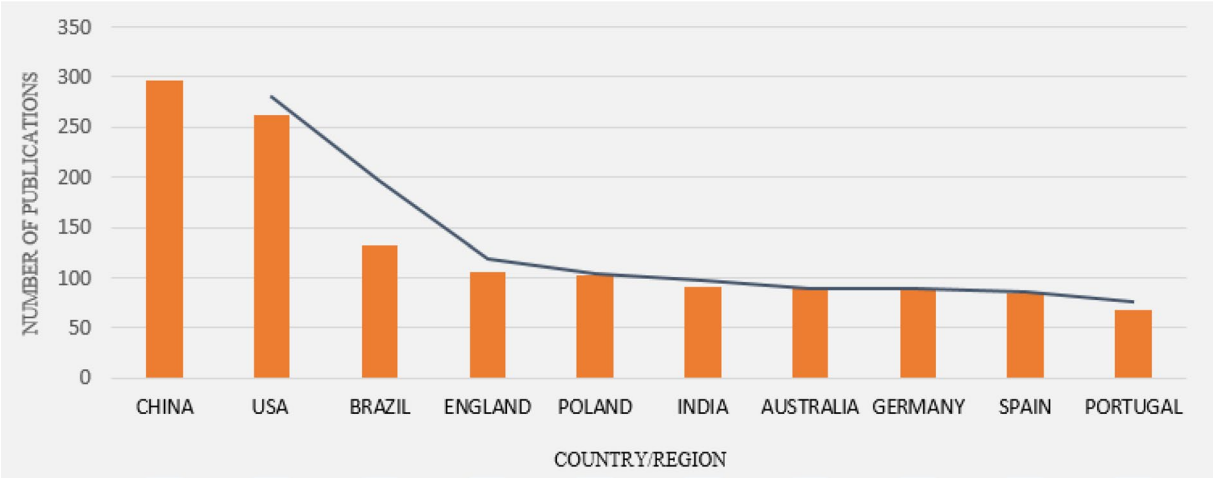

Fig. 3 The publications of the head 10 countries/regions

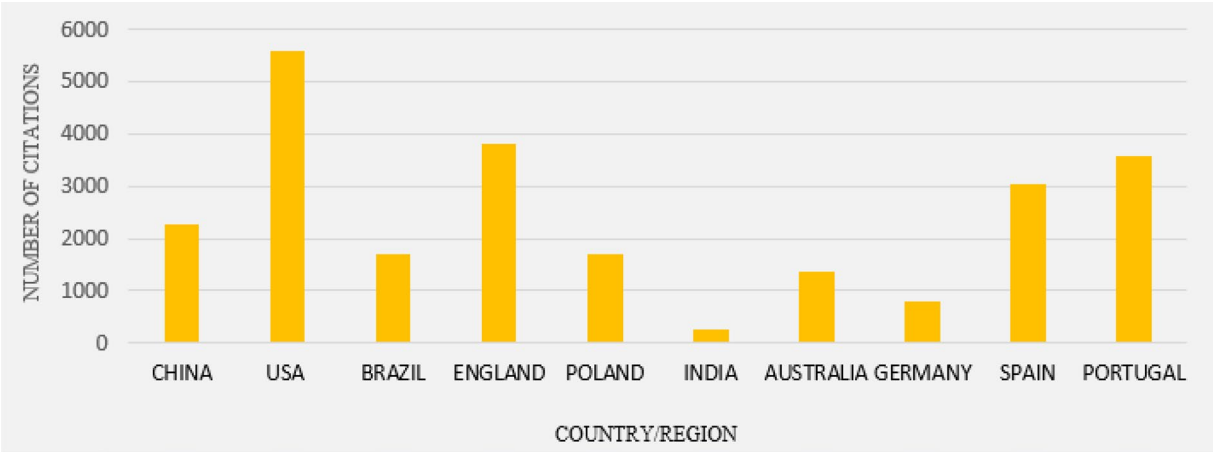

Fig. 4 The citations of the publications for the head 10 countries/regions

Table 3 The 10 most prolific institutions

Rank	Institution	Country/region	TP	TC	TC/TP	%TP	H-index
1	Universidade Do Porto	Portugal	49	3,314	368.2	3.13	19
2	University Of Technology Sydney	Australia	30	443	14.77	1.92	11
3	Wroclaw University Technology	Poland	28	301	10.75	1.8	8
4	Nanyang Technological University	Singapore	24	491	20.46	1.53	10
5	Universidade de Sao Paulo	Brazil	24	240	10	1.53	8
6	Universidade Federal de Pernambuco	Brazil	23	456	19.83	1.47	13
7	Wroclaw University of Science and Technology	Poland	23	594	25.83	1.47	7
8	Politecnico di Milano	Italy	21	543	25.85	1.34	8
9	Rowan University	USA	20	1,147	57.35	1.28	9
10	University of Waikato	New Zealand	19	558	29.37	1.21	8

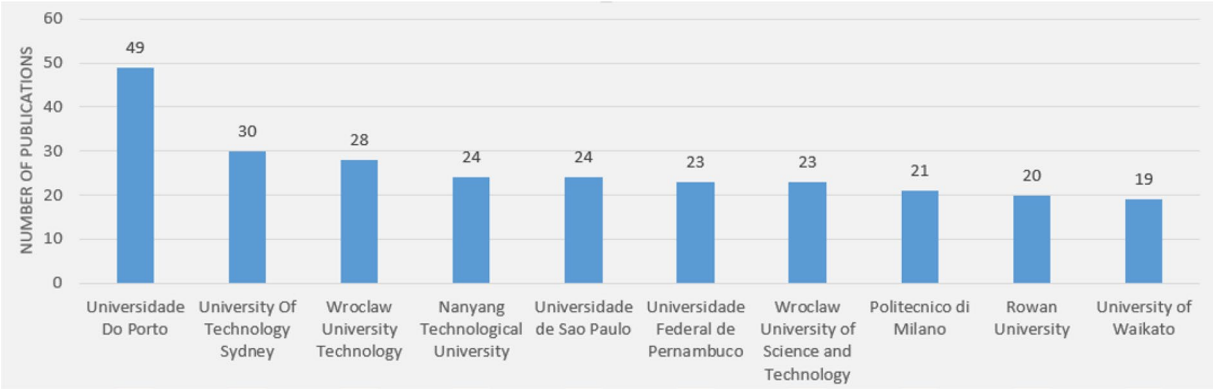

Fig. 5 The total number of publications for the head 10 institutions

University has a fewer number of publications than the University of Technology Sydney, it has a higher ratio of TC/TP.

Similar to that, Universidade Federal de Pernambuco yielded a greater proportion than Universidade De Sao Paulo, University of Technology Sydney, and the Wroclaw University of Science Technology; which indicates that the Universidade Federal de Pernambuco has a rising popularity while the others have a strong interest in concept drift. As regards, Portugal, Brazil, and the USA have the leading institutions in this specific research area.

The assessment of institutions inferred that Portugal, Australia, Poland, Singapore, Brazil, Italy, the USA, and New Zealand are passionate about concept drift-related research, respectively. One of the 10 institutions is located in Portugal, and two of them are located in Poland. Additionally, two *highly authoritative publications* are affiliated with the same institution called Universidade do Porto, which falls into the first place in the head 10 institutions list. The most fruitful two institutions are addressed in Brazil, and the others are located in Australia, Singapore, Italy, the USA, and New Zealand.

3.4 Prolific authors

The prolific authors could be stated as standard-bearers and tastemakers of scientific research outputs such as academic papers, conference papers, technical patents, etc. These authors lead the way for hot topics and trends for future research activities.

The analyses are conducted to exhibit the top 10 most prolific authors, in 3068 authors, regarding the publications. As a result, the most productive authors are represented in Table 4. Even if the list of productive authors is ranked according to the number of papers written, the TC/TP ratio is equally important. Gama J., who ranked first, has not only the highest number of publications in total but also the topmost h-index. On the other hand, Bifet A. performed the best score on TC/TP (89.14) with the second-highest h-index, which points out the rising fame per publication. Zliobaite I. pursues Bifet A. with a ratio TC/TP (85.28), whereas the author has only 18 papers. The third best ratio TC/TP (80.34) belongs to Gama J., which stands for immense recognition per publication. Therefore, the mentioned authors might be regarded as standard-bearers of concept drift.

Table 4 The 10 most prolific authors

Rank	Authors	Country/region	TP	TC	TC/TP	%TP	H-index
1	GAMA J	Portugal	42	3,374	80.34	2.69	20
2	WOZNIAK M	Poland	39	824	21.13	2.5	10
3	BIFET A	Spain	29	2585	89.14	1.85	13
4	KRAWCZYK B	USA	24	794	33.08	1.53	10
5	LU J	Australia	24	361	15.04	1.53	9
6	ZHANG GQ	Australia	24	402	16.75	1.53	10
7	POLIKAR R	USA	20	1147	57.35	1.28	9
8	MINKU LL	England	18	1,009	56.06	1.15	9
9	ZLIOBAITE I	Finland	18	1,535	85.28	1.15	11
10	KHAN L	USA	17	505	29.71	1.09	10

The feature analysis of the prevailing data represented that; 3 out of 10 prolific authors (Krawczyk B., Polikar R., Khan L.) are from the USA with similar rates of TC/TP. As depicted in the list, Lu J. and Zhang GQ. are the most fertile authors from Australia. The only author placed on the list from Poland is Wozniak M., who follows Gama J. closely with the number of publications. Minku LL. from England has a greater ratio of TC/TP than most prolific authors. Gama J. from Portugal, Bifet A. from Spain, and Zliobaite I. from Finland are the most rewarding authors who have a high incidence of citations. The total number of publications and citations of the most productive ten authors are illustrated in Fig. 6.

4 Cooperation networks of concept drift

The collaboration relationships of countries, institutions, and authors are illustrated with cooperation networks, and vigorous analyses are performed to indicate the improvement of concept drift regarding countries/regions, institutions, or authors.

4.1 Collaboration network for countries/regions

A collaboration network of countries/regions is created by VOS viewer, which is an outstanding tool for constructing bibliometric networks, to visualize the affiliation among the associated countries/regions from 1980 to 2020, shown in Fig. 7. The analysis rendered a result that the 51 countries/regions are separated into 10 particular clusters where each color symbolizes a different cluster. The head 10 prolific countries/regions in the collaboration network are China, the USA, Brazil, England, Poland, India, Australia, Germany, Spain, and Portugal, respectively.

The size of the nodes represents the overall number of publications after data cleansing, while the links, which connect diverse countries/regions, point out the collaborative relationship between the two. Provided that the links get darker and thicker, it means the number of collaborations is mounting among the two connected nodes. According to the collaboration networks, the USA is the most cooperative country and frequently collaborates with China and Poland. The thick and close links demonstrate that; there is an intense relationship between China and the USA, likewise Brazil and England or Spain and Portugal.

In related work, the three indicators are broadly used to characterize the cooperative countries/regions, which are; (a) Local links—the number of countries/regions that cooperated with the target country/region, (b) Link strength—the frequency of occurrence of collaborations, and (c) Cluster—the cluster in which the node is included. Detailed information about these three indicators is provided in Table 5 for the most prolific ten countries/regions. Not only the highest number of local links is associated with the USA (33), but also the best score of link strength (129) is.

Even though the most productive country/region is China, the USA cooperates with other target countries/regions more often. Although Brazil and Poland have a large number of publications, these countries' link strengths are comparatively low; meaning that the majority of the documents are published independently.

The local links and link strength value of India are equal to one, which means there is only one publication that has cooperated, which is with New Zealand. The same conditions are valid for one more country/region; Taiwan. Therefore, it is not possible to evaluate these countries as cooperative.

The most productive ten countries/regions in the country collaboration network are separated into 10 original clusters. Particularly, the USA is the part of Cluster 2; Poland and Spain are the elements of Cluster 3; Australia and Portugal are the parts of Cluster 4; England is in Cluster 5; Germany is a member of Cluster 6; Brazil is a staff of Cluster 7;

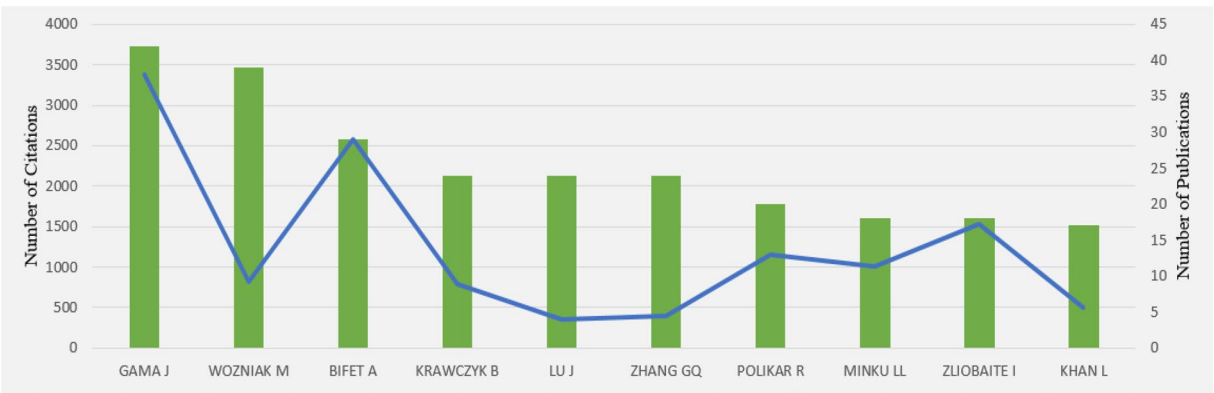

Fig. 6 The total number of publications and the citations for the head 10 authors

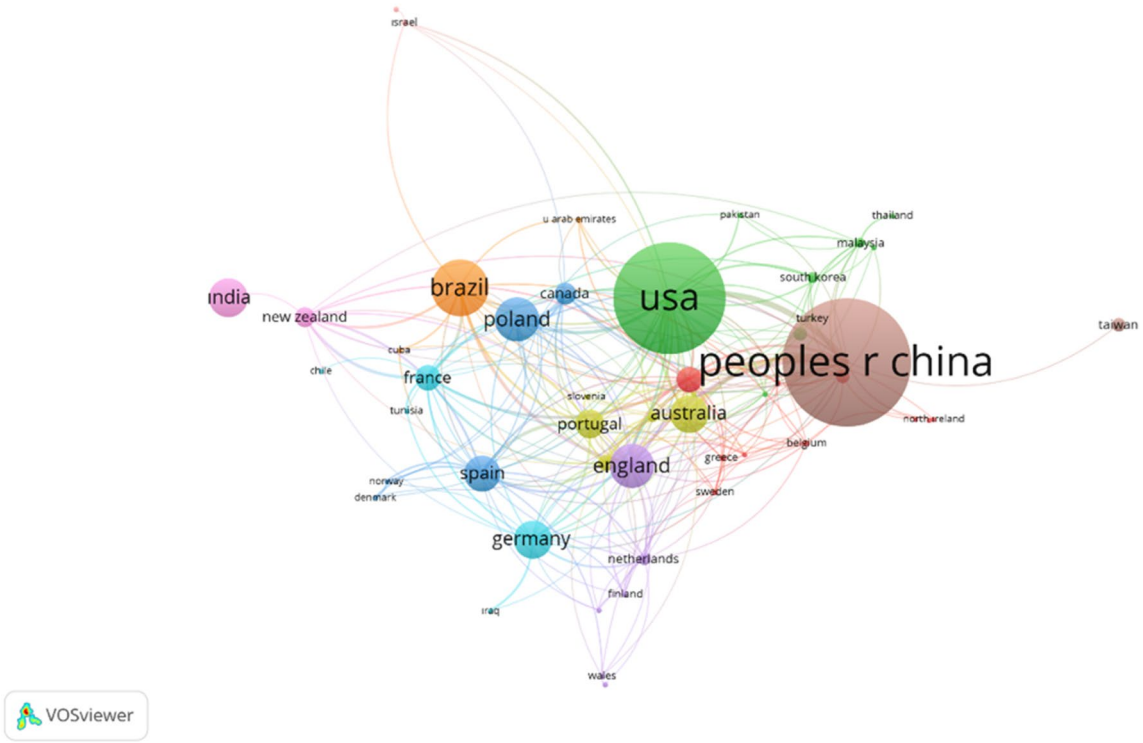

Fig. 7 The collaboration network of countries/regions from 1980 to 2020

Peoples are China joined Cluster number 8 and India is in Cluster 9.

4.2 Collaboration network at the level of institution

The collaboration network for institutions from 1980 to 2020 is constructed by a VOS viewer and demonstrated in Fig. 8. Throughout the analysis, 79 institutions of publications fall into 12 categories.

Table 5 The top 10 prolific countries/regions in the collaboration network

Rank	Target country/region	TP	Local links	Link strength	Cluster
1	Peoples R China	297	26	128	8
2	USA	262	33	129	2
3	Brazil	132	17	60	7
4	England	105	25	88	5
5	Poland	103	13	47	3
6	India	91	1	1	9
7	Australia	89	16	60	4
8	Germany	89	19	50	6
9	Spain	84	20	61	3
10	Portugal	68	18	58	4

In order to investigate the relationships between institutions, the top 10 prolific institutions in the collaboration network with three indicators, are listed in Table 6. Universidade do Porto is the superior institution, concerning both local links (19) and the link strength (32), followed by the University of Waikato with local links (14) and link strength (26). The link strengths of the Rowan University and Universidade Federal de Pernambuco are relatively low, which expresses that a remarkable amount of the related works are published individually.

The leading institutions are Universidade do Porto (Portugal), the University of Waikato (New Zealand), and Nanyang Technological University (Singapore), which exhibit the most cooperative behaviors among the head 10 institutions.

4.3 Collaboration network at the level of authors

The author collaboration network is constructed and visualized by VOS viewer software and displayed in Fig. 9, to clarify the relationships of authors. During the analysis, 109 authors, who have the corresponding publications, are inspected to reveal the collaborative relationships between authors. The top 10 prolific authors in the collaboration network, are listed in Table 7.

In total, 19 clusters are defined with the clustering method in VOS viewer, where each node represents an

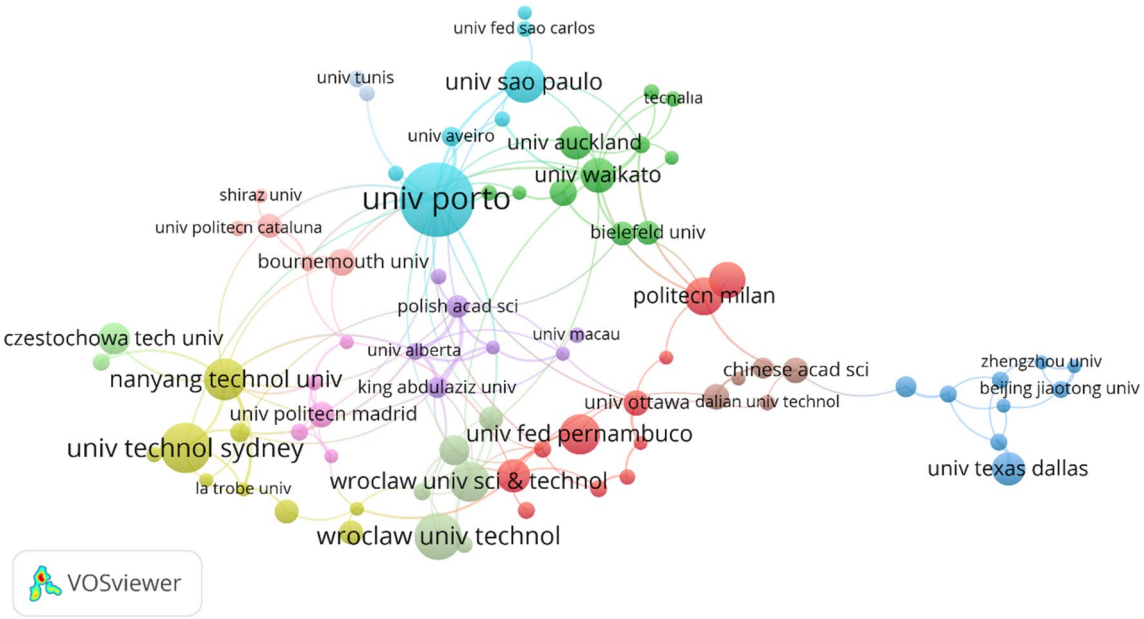

Fig. 8 The institution collaboration network from 1980 to 2020

Table 6 The top 10 most prolific institutions in the collaboration network

Rank	Institution	TP	Local links	Link strength	Cluster
1	Universidade Do Porto	49	19	32	6
2	University Of Technology Sydney	30	6	10	4
3	Wroclaw University Technology	28	3	5	7
4	Nanyang Technological University	24	11	22	4
5	Universidade de Sao Paulo	24	6	8	6
6	Universidade Federal de Pernambuco	23	3	4	1
7	Wroclaw University of Science and Technology	23	8	16	7
8	Politecnico di Milano	21	6	9	1
9	Rowan University	20	1	1	1
10	University of Waikato	19	14	26	2

author with a divergent color. Local links refer to the lines that connect one author to another, while the thickness and the distance of the lines indicate the strength of the relationship between authors. Gama J is the leading author in local links (27), followed by Bifet A (23); meaning that they are the most synergic authors. Nevertheless, Bifet A ranks first concerning the link strength (65), proving that the author's cooperation frequency is higher than others.

Even though the numbers of local links of Lu J (9) and Zhang GQ (9) are comparably small, their link strengths are (52 and 53) higher than Gama J. The number of local links of Wozniak M (12), who is another fruitful author, is higher than most of the authors on the list. Although Polikar R is a productive author, who has 20 unique related studies, both the number of local links (4) and the link strength (12) are far smaller than others.

The number of citations for an author is as important as the number of publications. The analysis of original data has proved that; 15,442 authors have been cited by a journal, a distinct number of times. Therefore, in addition to the number of publications, the citations of an author are used for analysis to determine influential authors. The minimum number of citations of an author is set to 49 for analysis, and 100 authors are on the threshold. A co-citation network of the authors in the customized time range has been represented in Fig. 10.

As a result, Gama J, Wozniak M, and Bifet A the most influential authors in concept drift-related research; indicating that they have contributed to concept drift intensely with a huge number of publications cited tons of times. Additionally, there are lots of precious authors, who have tremendous numbers of citations on concept drift research field, out of

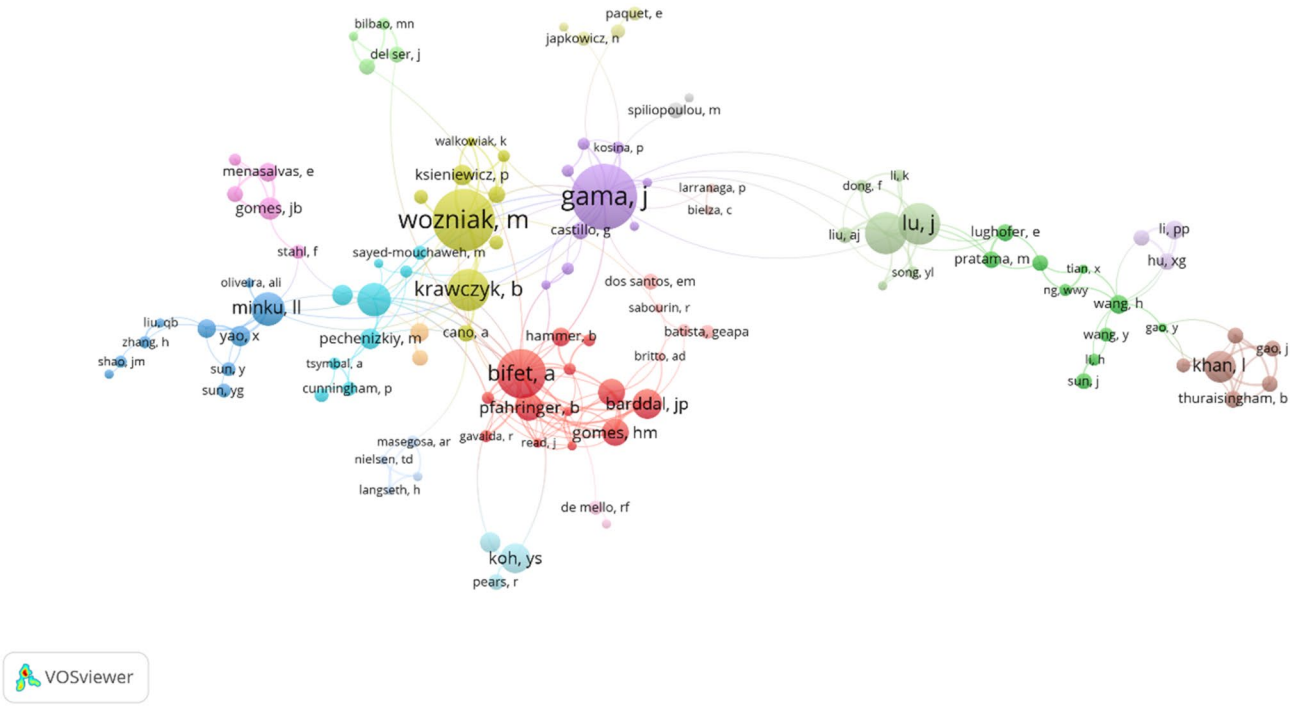

Fig. 9 The author's collaboration network from 1980 to 2020

Table 7 The top 10 prolific authors in the author collaboration network

Rank	Authors	Country/region	TP	Local links	Link strength	Cluster
1	Gama J	Portugal	42	27	50	4
2	Wozniak M	Poland	39	12	49	7
3	Bifet A	Spain	29	23	65	2
4	Krawczyk B	USA	24	8	25	7
5	Lu J	Australia	24	9	53	9
6	Zhang GQ	Australia	24	9	52	9
7	Polikar R	USA	20	4	12	5
8	Minku LL	England	18	9	20	6
9	Zliobaite I	Finland	18	9	19	8
10	Khan L	USA	17	6	34	1

the top 10 prolific authors list, such as; Widmer G, Domingos P, Demsar J, Hulten G, Wang H, Kolter JZ, Ditzler G, Brzezinski D, Baena-Garcia M, Tsymbal A, etc.

5 In-depth analysis of concept drift

The overall analyses of concept drift-related publications, concerned with the number and type of the papers, the number of citations, and the cooperation of networks, are conducted to disclose emerging areas in this specific field. Yet, highly authoritative publications, most prolific institutions, and influential authors have been declared.

Further, a deep analysis of 1564 publications is purposed, to exhibit the current challenges of this research area, trending/hot topics, and the future research directions. To explore hidden relationships of publications, certain types of analyses have been performed, which are; co-occurrence analysis, burst detection analysis, timeline view analysis, and bibliographic coupling analysis.

5.1 Burst detection analyses for authors, journals, and references

A citation burst implies that the scientific community has been fascinated by an incredible degree of attention towards the underlying contribution. The detection of a burst event,

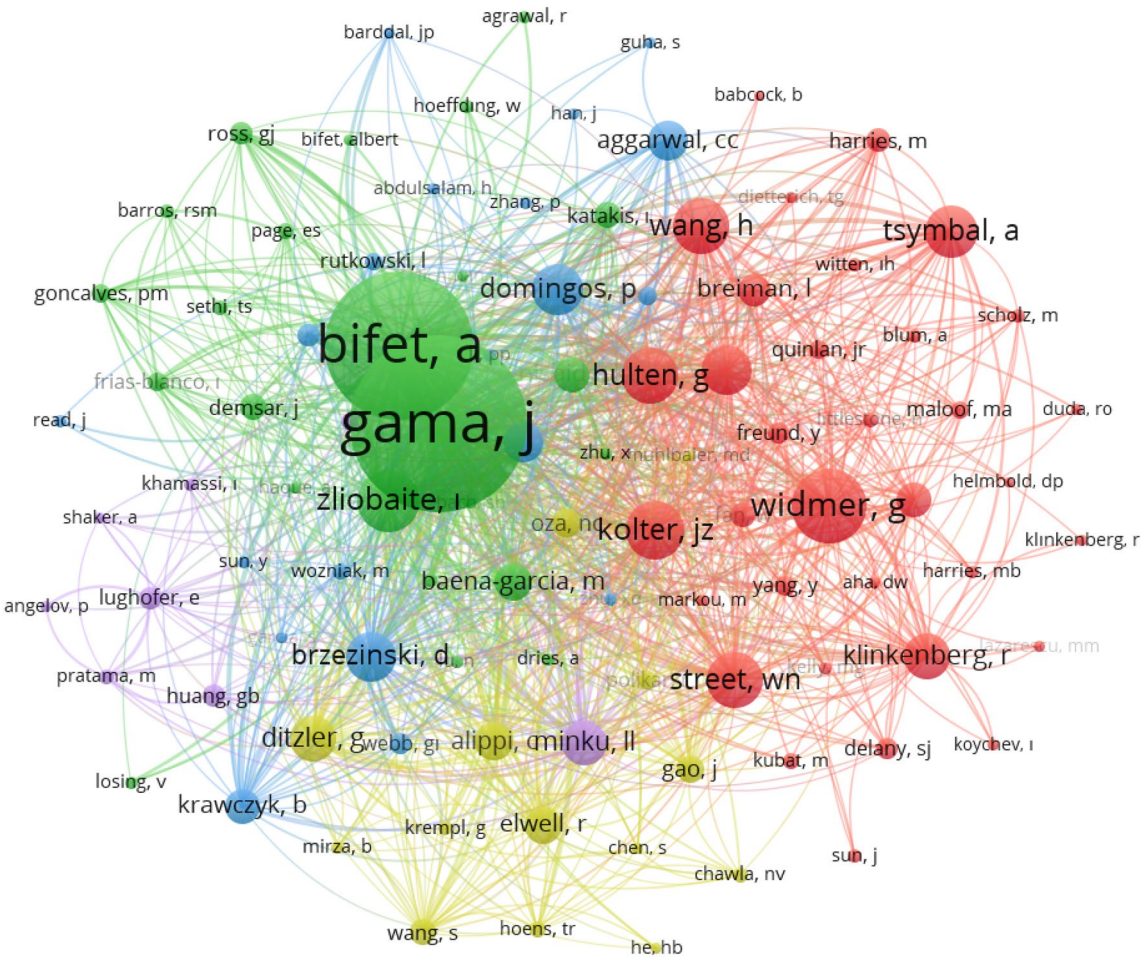

Fig. 10 The co-citation network of the authors

which could be the most prepotent area in the research field, could last several years as well as a year. Moreover, if a cluster consists of plentiful nodes with intense citation bursts; the entire cluster snatches a hot area of research or an emerging trend. The Cite Space software employs Kleinberg's algorithm (Kleinberg and Tardos 1999) during burst detection.

The burst detection analysis displays the vigorous changes in concept drift-related publications. In this section, analyses are individually performed to detect bursts and the lasting period of bursts for authors, journals, and references. From 1980 to 2020, the head 12 authors with the strongest citation bursts are listed in Table 8. Regarding that, Gama J has the greatest strength (134.36), while the longest citation burst duration belongs to Widmer G with 13 years from 2004 to 2017.

In addition, Bifet A has the second-highest strength with 104.62, within 7 years from 2013 to 2020. Although the top 12 cited authors' strengths are quite satisfactory, their reputation might diverge over time, related to the citation burst periods. Moreover, the citation bursts of four authors, i.e., Gama J, Bifet A, Zliobaite I, and Ditzler G, are the closest to the present. Considering the burst ending time; the oldest citation burst, among the top 12 authors, belongs to Domingos P with a strength of 15.75.

Similar to cited authors, the head 12 journals with the strongest citation bursts are represented in Table 9, meaning that these head journals in the table have been cited repetitiously in a certain period of time. Through the analysis results, the ACM Computing Surveys journal has the highest strength value (64.96), followed by KDD '03: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (59.96) and Machine Learning (39.96).

The ACM Computing Surveys journal desires to improve perspectives and recognize trends in complex technologies. The expected contribution from publications is to bridge existing and emerging technologies with a variety of science and engineering domains. On the other hand, KDD '03: Proceedings of the Ninth ACM SIGKDD International

Table 8 The head 12 cited authors with the strongest citation bursts from 1980 to 2020

Rank	Cited Authors	Strength	Begin	End	1980–2019
1	Gama J	134.36	2015	2020	
2	Bifet A	104.62	2013	2020	
3	Widmer G	79.78	2004	2017	
4	Wang H	67	2010	2017	
5	Zliobaite I	58.05	2016	2020	
6	Ditzler G	56.59	2018	2020	
7	Street WN	40.22	2015	2018	
8	Kolter JZ	37.33	2009	2013	
9	Klinkenberg	36.73	2003	2011	
10	Hulten G	36.7	2002	2011	
11	Tysmbal A	26.39	2008	2011	
12	Domingos P	15.75	2002	2008	

Conference on Knowledge Discovery and Data Mining consist of invaluable publications since it is the original international platform for data mining researchers.

Nevertheless, the journal of Machine Learning publishes articles that have quite essential outputs on a broad range of learning methods, and learning from drifting dynamic data has become an urgent topic. The citation burst of the cited journal of KDD '03: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining holds the longest duration of 12 years from 2005 to 2017.

As first treaters, several journals started earlier to deal with concept drift (i.e., from 2003) which are Machine Learning, KDD '07: Proceedings of the Seventh ACM International Conference on Knowledge Discovery and Data Mining, and KDD '03: Proceedings of the Ninth ACM International Conference on Knowledge Discovery and Data Mining. Some cited journals' dominant citation bursts have persisted to 2020, i.e., ACM Computing Surveys, Neurocomputing, Information Sciences, IEEE Transactions on Neural Networks and Learning Systems, Pattern Recognition, and Expert Systems with Applications. It's proved that these six influential journals have a tastemaker impact on concept drift-related publications.

An identical burst detection analysis is conducted with Cite Space, from 1980 to 2020, to ascertain the references. There are a total of 21,531 valid references related to concept drift. In Table 10, the head 12 references with the strongest citation bursts from 1980 to 2020, are listed. Gama J, 2014, ACM Computing Surveys, V46; ranks first with the highest strength (156.6). Even the second-highest value of strength (53.2) is much less; which proves that Gama J, 2014, ACM Computing Surveys, V46 is the most well-known citation in this field.

Additionally, the cited reference of Gama J, 2014, ACM Computing Surveys also has the longest duration, which is 5 years, from 2015 to 2020. The strongest citation bursts of five cited references, i.e., Gama J, 2014, ACM Computing

Table 9 The head 12 cited journals with the strongest citation bursts from 1980 to 2020

Rank	Cited Journals	Strength	Begin	End	1980–2019
1	ACM Computing Surveys	64.96	2016	2020	
2	KDD '03	59.96	2005	2017	
3	Machine Learning	39.96	2003	2013	
4	Intelligent Data Analysis	16.03	2006	2014	
5	Neurocomputing	34.22	2015	2020	
6	KDD '07	32.92	2003	2011	
7	Information Sciences	31.16	2017	2020	
8	IEEE Trans. Neur. Learn	29.30	2017	2020	
9	Intell Data Analysis	26.92	2009	2016	
10	Pattern Recognition	25.03	2018	2020	
11	KDD '09	24.92	2011	2014	
12	Exp. Sys. with App	22.4	2018	2020	

Table 10 The head 12 cited references with the strongest citation burst from 1980 to 2020

Rank	Cited References	Strength	Begin	End	1980–2019
1	Gama J, 2014, ACM Computing Surveys, 46	156.6	2015	2020	
2	Ditzler G, 2015, IEEE Comput. Intelligence Magazine, 10, 12	53.2	2017	2020	
3	Brzezinski D, 2014, IEEE Trans. on Neur. Net. and Lear. Sys., 25, 81	42.49	2016	2020	
4	Krawczyk B, 2017, Information Fusion, 37, 132	40.72	2018	2020	
5	Elwell R, 2011, IEEE Transactions on Neural Networks, V22, P1517	36.18	2013	2016	
6	Frias-Blanco I, 2015, IEEE Trans on Know and Data Eng, V27, P810	29.82	2018	2020	
7	Minku LL, 2012, IEEE Trans. on Know and Data Eng, V24, P619	24.77	2014	2017	
8	Bifet A, 2010, Jour. of Mac.Lear. Research, V11, P1601	24.53	2013	2015	
9	Kolter JZ, 2007, Journal of Machine Learning Research, 8, 2755	20.95	2009	2012	
10	Minku LL, 2012, IEEE Trans. on Knowl. and Data Eng., 22, 730	19.9	2011	2015	
11	Bifet A, 2009, KDD '09: Proc. of the 15th Int. Con. on Know. disc and data min, 0, 139	19.4	2011	2014	
12	Klinkenberg R, 2004, Intel. Data Anal., 8, 281	13.25	2007	2009	

Surveys, V46, Ditzler G, 2015, IEEE Comput. Intelligence Magazine, V10, P12, Brzezinski D, 2014, IEEE Trans. on Neur. Net. and Lear. Sys., V25, P81, Krawczyk B, 2017, Information Fusion, V37, P132, and Frias-Blanco I, 2015, IEEE Trans on Know and Data Eng, V27, P810, have lasted to 2020, which are closest to the present.

5.2 Timeline view and co-occurrence analysis of keywords

Text-mining, the examination of keywords, may help to capture the main theme of research, which is crucial for comprehending the trendy topics in the research field. Co-occurrence analysis, which is in fact the counting of paired data, is a quantitative analysis method and is a compelling tool to encourage data mining and knowledge discovery.

As a result of the analysis conducted; there are 2560 keywords in total. For each keyword; the minimum number of occurrences is set as 10, and 59 keywords satisfy the threshold. The keywords are further separated into 10 clusters. The co-occurrence network of keywords is visualized by the VOS viewer and represented in Fig. 11.

The thickness of the nodes represents the frequency of that keyword in publications and the local links connect nodes refer to the relationships between keywords. A certain number of keywords occur repeatedly, such as "concept drift", "data stream", "machine learning", "classification", "online learning" and "data stream mining". Therefore, a great proportion of concept drift-related publications targets handling concept drift in streaming data with online learning algorithms or machine learning.

In addition to the co-occurrence of keywords analysis, a timeline view analysis of keywords is supervised, to capture the trends of hot topics in concept drift. The process is handled with the service of Cite Space; the keywords are divided into 8 clusters and the outcomes are represented in Fig. 12. The largest cluster is ''concept drift", followed by the second-largest cluster which is "data stream", which means that the concept drift in data streams is a critical problem and has been extensively investigated by authors. The third cluster is "recommendation system", which highlights the most popular area of concept drift-related publications.

The rest of the clusters in order are; ''passive concept drift", ''dynamic financial distress prediction", ''learning approaches", ''learning evaluation", ''eccentricity data analytics", "streaming data", and "predictor weight". The continuous change in the clusters proves that publications evolve from time to time. The existence of eccentricity data analytics in streaming data is a novel approach, likewise, the use of predictor weights. The timeline view of keywords demonstrated the research hotspots in the customized time range, namely from 1980 to 2020. Researchers, who are

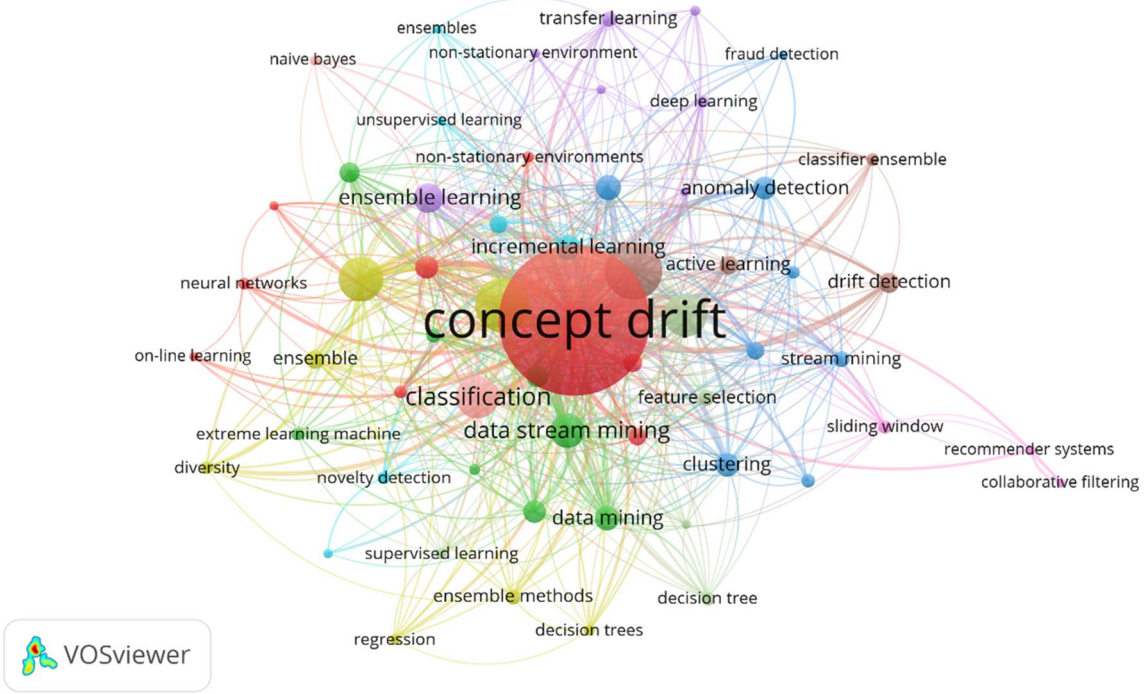

Fig. 11 A co-occurrence network of keywords related to concept drift

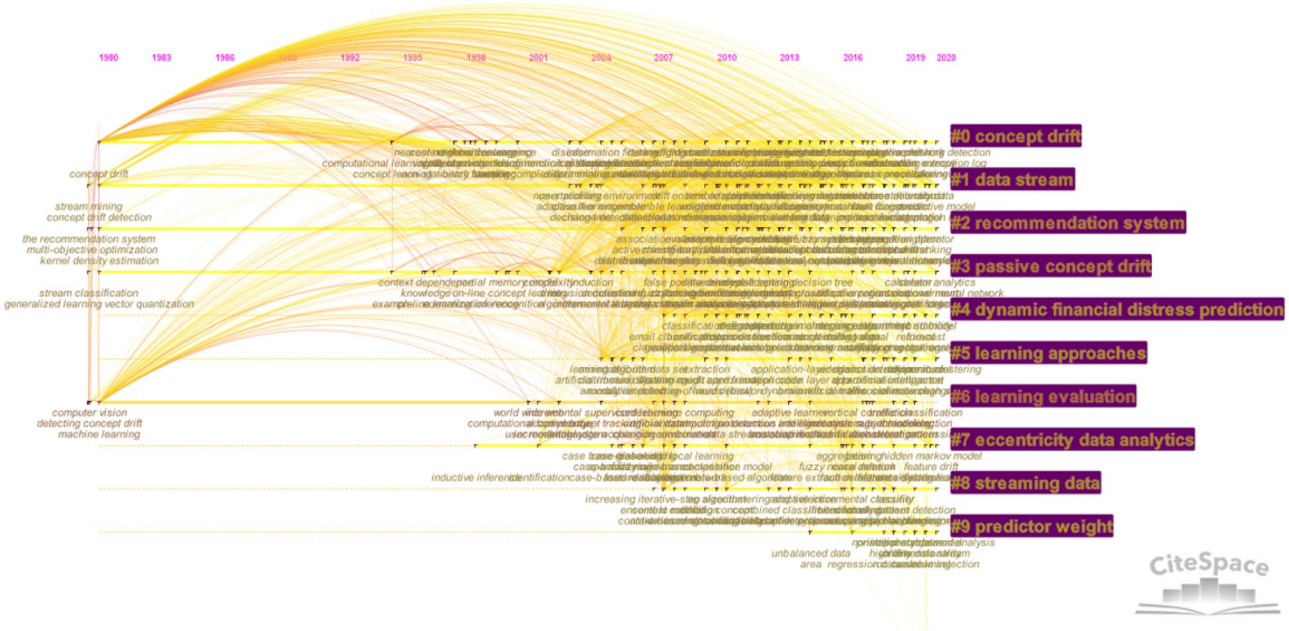

Fig. 12 The timeline view of keywords in concept drift

interested in the concept drift area, might benefit from this analysis for further studies.

5.3 Bibliographic coupling analysis

Bibliographic coupling is a phenomenon that occurs if the same paper(s) is cited in two articles, and refers to the

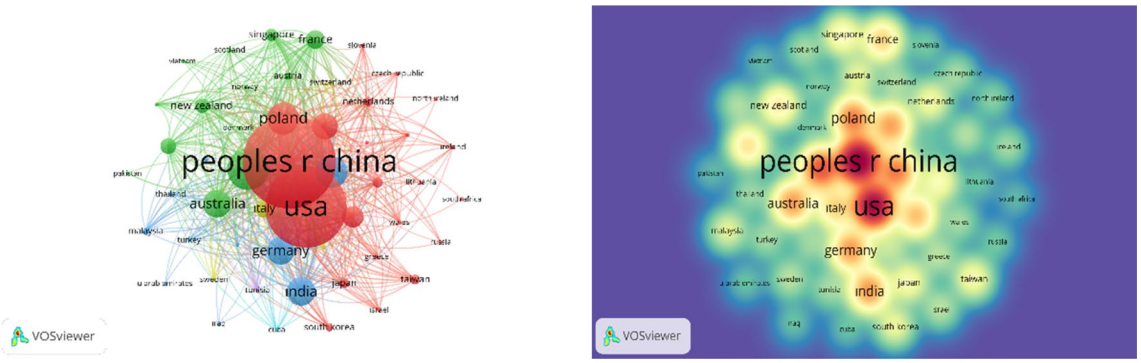

Fig. 13 The visualization of the network and the density of countries/regions

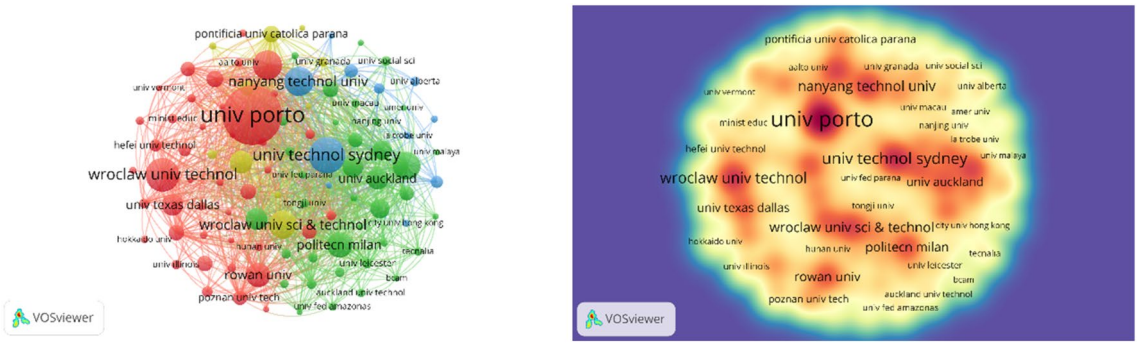

Fig. 14 The visualization of the network and the density of institutions

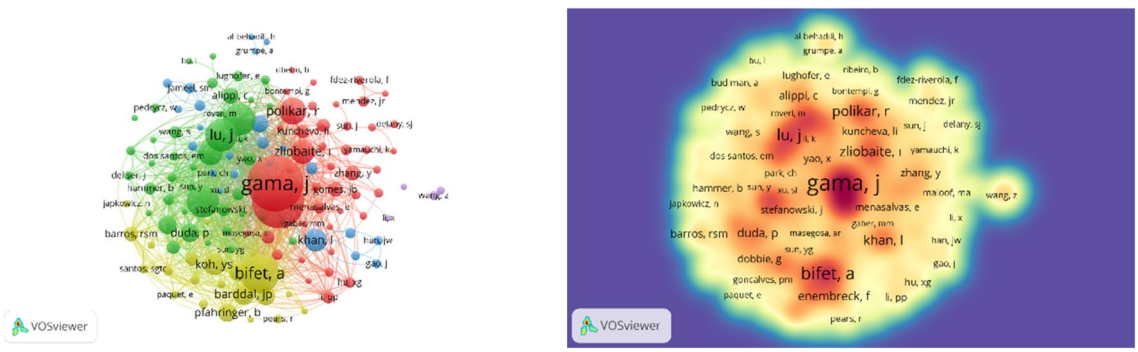

Fig. 15 The visualization of the network and the density of authors

relationship between the two publications. The coupling implies the similarity of the subject argument of the two publications, which helps the researcher to find the related work in the past. The higher the coupling degree is, the closer the subject matter of the two publications gets.

The analysis has been performed with a VOS viewer, to examine the degree of coupling for disparate levels, such as; the level of countries or regions, the level of authors, and the level of institutions. According to the results of these analyses, from 1980 to 2020, the network and the density visualizations of every level are shown in Figs 13, 14 and 15, respectively.

In the network visualization graphs, every color defines a divergent cluster, and the nodes' size illustrates the degree of coupling. According to the bibliographic coupling of countries/regions; China, with the highest coupling degree, interacts closely with the USA, Poland, Spain, Portugal, Brazil, Germany, and Australia. Simultaneously, Brazil has tight relationships with England, France, India, New Zealand, and Singapore.

As a consequence of the bibliographic coupling analysis of institutions, the highest-ranked four institutions, one from each cluster, are demonstrated as the outstanding institutions, which are Universidade Do Porto, University Of Technology Sydney, Universidade Federal de Pernambuco, and Wroclaw University of Science and Technology.

Regarding the bibliographic coupling analysis, Gama J, Bifet A, Polikar R, and Lu J have the greatest degree of coupling, implying that scholars who do research in the field of concept drift favor citing publications belonging to these authors. Furthermore, the bibliographic coupling analysis of authors reveals the unbiased relation among authors.

6 Conclusion

In this paper, a large-scale overview and visualized analysis of 1,564 publications related to *concept drift* from 1980 to 2020, which are obtained from the Web of Science database, is represented. Two disparate skillful bibliometric tools, i.e., VOS viewer and Cite Space, are employed to observe the research clusters, current status, and relationships between clusters. The primary objective of this study is to offer insight into the possible future directions for more valuable further research.

Through the overall analysis of valid literature data, fundamental bibliometric indicators are explored, in the manner of three levels, based on the publications, citations, and cooperation of networks, quantitatively. During the analyses, the total number of publications, and citations, the average number of citations per publication, h-index, and the relationships of nodes in the networks (local links, and link strengths) are evaluated to shed light on the popular research directions, most prolific countries/regions, highly authoritative publications, most effective authors, generative institutions, and the progression of the publications' citations. In consideration of the analyses results, a brief conclusion of the principal findings of this study can be summarized as;

(1) Proceedings papers and articles occupy a relatively huge amount of all papers, and a greater proportion of research harvests are based upon the theoretical foundation, such as computer science, engineering, etc.
(2) In the last four decades, China is the most prolific country with the highest number of publications. Whereas, the USA is the most cooperative country/region with the highest h-index. Portugal has the highest citation per publication.
(3) The most fruitful institution is revealed as the Universidade do Porto, which has the highest number of publications, and h-index. According to the institution collaboration network, Universidade do Porto is the superior institution exhibiting the most cooperative behaviors.
(4) Gama J, Portugal, is the most prolific author, with the topmost h-index. Nevertheless, Bifet A, Spain, has the highest recognition per publication; followed by Zliobaite I from Finland. Considering the author's collaboration network, Gama J, Wozniak M, and Bifet A have the most cooperated publications related to concept drift.

In-depth analyses, such as; burst detection analysis, timeline view analysis, co-occurrence analysis, and bibliographic coupling analysis, are performed specifically to diagnose the current challenges, emerging trends, and possible future directions. According to these analyses, the key findings of the concept drift-related publications are defined: I) the visualization of the co-occurrence network of keywords reveals that the research area of concept drift is still restricted since the one and largest node denotes "concept drift".

Additionally, the keywords present a map of trends for researchers. For instance, the popular methods of handling concept drift are gathered in the clusters of online/incremental learning, ensemble learning, supervised/unsupervised learning, deep learning, active learning, transfer learning, extreme learning machine, drift detection methods, and neural networks, which are the most used approaches. In fact, unsupervised learning and semi-supervised learning have gained more attention lately over supervised learning since the label for each concept is not available in streaming data.

The drift detection methods might be categorized into three; ensemble learning, windowing technique, and statistical process control. Plentiful papers could be found in the literature that relies on SPC to detect drifts, which trace the online error rate evolution of base learners. While the significance test level is exceeded, the concept drift is assumed to have occurred which leads to the exchange of base learners. The very first and widely-used drift detector is DDM, which considers Binomial distribution, followed by Early Drift Detection Method (EDDM) (Baena-Garcia et al. 2006), and Reactive Drift Detection Method (RDDM) (Barros et al. 2017). Hoeffding Drift Detection Method (HDDM) (Frías-Blanco et al. 2015) modifies DDM by employing Hoeffding's inequality, with two variants; $HDDM_A$ for abrupt drifts and $HDDM_W$ for gradual drifts. Fast Hoeffding Drift Detection Method (FHDDM) (Pesaranghader and Viktor 2016) focuses on the drawbacks of HDDM, followed by Stacking Fast Hoeffding Drift Detection Methos (FHDDMS) and Additive ($FHDDMS_{add}$). Page-Hinkley test (PHT) is based on the PH statistics, whereas Spectral Entropy Drift Detector (SEDD) (Chikushi et al. 2020) computes the spectral entropy along the error stream. EWMA for Concept Drift Detection (ECDD) employs weighted moving average charts, which is a well-known algorithm to detect concept drifts.

Window-based detectors divide the data stream into windows based on data size or time interval in a sliding manner.

These methods monitor the performance of the most recent observations introduced to the learner and compare it with the performance of a reference window. While Adaptive Windowing (ADWIN) (Bifet and Gavaldà 2007) is the most popular window-based detector, SEED, and STEPD are the evolved forms of this type of detector. Wilcoxon Rank Sum Test Drift Detector (WSTD) (Barros et al. 2018) is inspired by STEPD, likewise Fisher Test Drift Detector (FTDD), Fisher Square Drift Detector (FSDD), and Fisher Proportions Drift Detector (FPDD) (Cabral and Barros 2018). Similarly, McDiarmid Drift Detection Method (MDDM) (Pesaranghader et al. 2018) slides a window over prediction results, which uses McDiarmid's inequality.

Ensemble learning methods employ multiple various base learners to operate. Most of the ensemble-based detectors are constructed on the Weighted Majority Algorithm (WMA) method. Streaming Ensemble Algorithm (SEA) (Street and Kim 2001)is one of the earliest methods to handle concept drift, followed by Accuracy Weighted Ensemble (AWE). Dynamic Weighted Majority (DWM) (Kolter and Maloof 2003) is the most popular passive ensemble method inspired by WMA, which evolved into Heterogeneous Dynamic Weighted Majority (HDWM) (Idrees et al. 2020) and Recurring Dynamic Weighted Majority (RDWM) (Sidhu and Bhatia 2019). These methods are only a few, there are many other algorithms built up to detect drifts accurately and precisely. Drift detection is quite important while dealing with evolving data. These methods have been used especially on recommender systems, fraud detection, evolving fuzzy systems, evolving neural networks, anomaly detection, feature selection, collaborative filtering, and novelty detection.

Many publications have represented similar classical methods to handle concept drift, such as classification and clustering. Even though online learning or machine learning (Bifet and Gavaldà 2009; Žliobaitė 2010; Loo and Marsono 2016) methods are frequently encountered, the current challenge is handled properly with the ensembles of detectors (Bifet et al. 2009; Maciel et al. 2015; Barros and Santos 2019), evolving fuzzy systems (Pratama et al. 2018), and artificial neural network algorithms (Jagait et al. 2021; Qiao et al. 2023) at a vast scale. Since the evolving systems adopt an open structure, where its components might be necessarily generated, pruned, merged, and recalled, are well suited to a given problem. II) Based on the timeline view analysis of keywords, the emerging trend/hot topics are revealed as cluster labels; *dynamic financial distress prediction*, *eccentricity data analytics*, and usage of *predictor weights*.

III) The possible future directions of further research are forecasted, correlated to the analysis of keywords, and pointed out as; partial memory learning or partial instance memory, learning vector quantization, word probability, p2p networks, face detection, and advanced computer vision systems. It'd be interesting that scholars prefer to propose a novel approach to solve the specific problems stated as future directions.

The extensive bibliometric analysis of the concept drift has been represented and the contributions of this paper are reviewed, which are crucial for researchers who are interested in concept drift. The findings assist the scholars while tracking the evolution of concept drift; proposing a novel method for handling concept drift, referring to the most prolific authors, investigating the latest algorithms used to detect concept drift, etc., from different view angles.

The drawback of this paper, which is required to be enhanced imminently, is; the limitation of the data source. Even though the Web of Science database hands over a mass of leading journals accessible with detailed information, is not capable of retaining the entire concept drift-related publications at last. In the future, the study is planned to be enlarged with divergent data sources and focused on concept drift analysis combined with text mining.

Funding The authors did not receive support from any organization for the submitted work.

Declarations

Conflict of interest All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Abdallah ZS, Gaber MM, Srinivasan B, Krishnaswamy S (2016) AnyNovel: detection of novel concepts in evolving data streams: an application for activity recognition. Evol Syst 7:73–93. https://doi.org/10.1007/s12530-016-9147-7

Abdullatif A, Masulli F, Rovetta S (2018) Clustering of nonstationary data streams: a survey of fuzzy partitional methods. Wiley Interdiscip Rev Data Min Knowl Discov. 8:e1258. https://doi.org/10.1002/widm.1258.

Agrahari S, Singh AK (2021) Concept drift detection in data stream mining : a literature review. J King Saud Univ Comput Inf Sci. 34:9523–9540. https://doi.org/10.1016/j.jksuci.2021.11.006

Alonso S, Cabrerizo FJ, Herrera-Viedma E, Herrera F (2009) h-Index: a review focused in its variants, computation and standardization for different scientific fields. J Informetr. 3:273–289. https://doi.org/10.1016/j.joi.2009.04.001

Anupama N, Jena S (2019) A novel approach using incremental oversampling for data stream mining. Evol Syst 10:351–362. https://doi.org/10.1007/s12530-018-9249-5

Babüroğlu ES, Durmuşoğlu A, Dereli T (2021) Novel hybrid pair recommendations based on a large-scale comparative study of concept drift detection. Expert Syst Appl. 163:1137. https://doi.org/10.1016/j.eswa.2020.113786

Baena-Garcia M, Campo-Avila J, Fidalgo R, et al (2006) Early drift detection method. In: 4th ECML PKDD international workshop on knowledge discovery from data streams

Barros RSM, Santos SGTC (2018) A large-scale comparison of concept drift detectors. Inf Sci (n Y). https://doi.org/10.1016/j.ins.2018.04.014

Barros RSM, Cabral DRL, Gonçalves PM, Santos SGTC (2017) RDDM: reactive drift detection method. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2017.08.023

Bayram F, Ahmed BS, Kassler A (2022) From concept drift to model degradation: an overview on performance-aware drift detectors. Knowl Based Syst. 245:108632. https://doi.org/10.1016/j.knosys.2022.108632

Bifet A, Holmes G, Kirkby R, Pfahringer B (2010) MOA: massive online analysis. J Mach Learn Res 11:1601–1604

Bifet A, Gavaldà R (2007) Learning from time-changing data with adaptive windowing. In: Proceedings of the 2007 SIAM international conference on data mining. https://doi.org/10.1137/1.9781611972771.42

Bifet A, Gavaldà R (2009) Adaptive learning from evolving data streams. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)

Bifet A, Holmes G, Pfahringer B, et al (2009) New ensemble methods for evolving data streams. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining - KDD '09

Bifet A, Hammer B, Schleif FM (2019) Recent trends in streaming data analysis, concept drift and analysis of dynamic data sets. ESANN 2019 - Proceedings, 27th European symposium on artificial neural networks, computational intelligence and machine learning 421–430

Borgman CL, Furner J (2002) Scholarly communication and bibliometrics. Annual Review of Information Science and Technology 36:2–72. https://doi.org/10.1002/aris.1440360102

Chikushi RTM, de Barros RSM, da Silva MGNM, Maciel BIF (2020) Using spectral entropy and bernoulli map to handle concept drift. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2020.114114

Cobo MJ, López-Herrera AG, Herrera-Viedma E, Herrera F (2011) Science mapping software tools: review, analysis, and cooperative study among tools. J Am Soc Inform Sci Technol. https://doi.org/10.1002/asi.21525

de Barros RSM, de Santos SGTC (2019) An overview and comprehensive comparison of ensembles for concept drift. Inf Fus 52:213–244. https://doi.org/10.1016/j.inffus.2019.03.006

de Cabral DR, de Barros RSM (2018) Concept drift detection based on fisher's exact test. Inf Sci (n Y). https://doi.org/10.1016/j.ins.2018.02.054

de Barros RSM, Hidalgo JIG, de Cabral DRL (2018) Wilcoxon rank sum test drift detector. Neurocomputing 275:1954–1963. https://doi.org/10.1016/j.neucom.2017.10.051

Ditzler G, Roveri M, Alippi C, Polikar R (2015) Learning in non-stationary environments: a survey. IEEE Comput Intell Mag 10:12–25. https://doi.org/10.1109/MCI.2015.2471196

Dong F, Zhang G, Lu J, Li K (2018) Fuzzy competence model drift detection for data-driven decision support systems. Knowl Based Syst 143:284–294. https://doi.org/10.1016/j.knosys.2017.08.018

Elwell R, Polikar R (2009) Incremental learning of variable rate concept drift. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) 5519 LNCS:142–151. https://doi.org/10.1007/978-3-642-02326-2_15

Elwell R, Polikar R (2011) Incremental learning of concept drift in nonstationary environments. IEEE Trans Neural Netw 22:1517–1531. https://doi.org/10.1109/TNN.2011.2160459

Falagas ME, Pitsouni EI, Malietzis GA, Pappas G (2008) Comparison of pubmed, scopus, web of science, and google scholar: strengths and weaknesses. FASEB J. https://doi.org/10.1096/fj.07-9492lsf

Frías-Blanco I, Del Campo-Ávila J, Ramos-Jiménez G et al (2015) Online and non-parametric drift detection methods based on Hoeffding's bounds. IEEE Trans Knowl Data Eng. https://doi.org/10.1109/TKDE.2014.2345382

Gama J, Medas P, Castillo G, Rodrigues P (2004) Learning with drift detection Advances in Artificial Intelligence - SBIA 2004, 17th Brazilian Symposium on Artificial Intelligence, São Luis, Maranhão, Brazil 3171: 286–29. https://doi.org/10.1007/978-3-540-28645-5_29

Gama J, Žliobaitė I, Bifet A et al (2014) A survey on concept drift adaptation. ACM Comput Surv. 46:1–37. https://doi.org/10.1145/2523813

Gemaque RN, Costa AFJ, Giusti R, dos Santos EM (2020) An overview of unsupervised drift detection methods. Wiley Interdiscip Rev Data Min Knowl Discov 10:e1381

Giusti L, Carvalho L, Gomes AT et al (2022) Analyzing flight delay prediction under concept drift. Evol Syst. https://doi.org/10.1007/s12530-021-09415-z

Gözüaçık Ö, Can F (2021) Concept learning using one-class classifiers for implicit drift detection in evolving data streams. Artif Intell Rev 54:3725–3747. https://doi.org/10.1007/s10462-020-09939-x

He X, Wu Y, Yu D, Merigó JM (2017) Exploring the ordered weighted averaging operator knowledge domain: a bibliometric analysis. Int J Intell Syst. https://doi.org/10.1002/int.21894

Henzgen S, Strickert M, Hüllermeier E (2014) Visualization of evolving fuzzy rule-based systems. Evol Syst 5:175–191. https://doi.org/10.1007/s12530-014-9110-4

Heusinger M, Raab C, Schleif FM (2022) Dimensionality reduction in the context of dynamic social media data streams. Evol Syst 13:387–401. https://doi.org/10.1007/s12530-021-09396-z

Hidalgo JIG, Maciel BIF, Barros RSM (2019) Experimenting with prequential variations for data stream learning evaluation. Comput Intell 35:670–692. https://doi.org/10.1111/coin.12208

Hoens TR, Polikar R, Chawla NV (2012) Learning from streaming data with concept drift and imbalance: an overview. Progress in Artificial Intelligence 1:89–101. https://doi.org/10.1007/s13748-011-0008-0

Hosseini MJ, Ahmadi Z, Beigy H (2013) Using a classifier pool in accuracy based tracking of recurring concepts in data stream classification. Evol Syst 4:43–60. https://doi.org/10.1007/s12530-012-9064-3

Hu H, Kantardzic M, Sethi TS (2019) No free lunch theorem for concept drift detection in streaming data classification : a review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10:e1327. https://doi.org/10.1002/widm.1327

Huang DTJ, Koh YS, Dobbie G, Pears R (2015) Detecting volatility shift in data streams. Proc IEEE Int Conf Data Min. https://doi.org/10.1109/ICDM.2014.50

Idrees MM, Minku LL, Stahl F, Badii A (2020) A heterogeneous online learning ensemble for non-stationary environments. Knowl Based Syst. https://doi.org/10.1016/j.knosys.2019.104983

Iwashita AS, Papa JP (2019) An overview on concept drift learning. IEEE Access 7:1532–1547. https://doi.org/10.1109/ACCESS.2018.2886026

Jagait RK, Fekri MN, Grolinger K, Mir S (2021) Load forecasting under concept drift: online ensemble learning with recurrent neural network and ARIMA. IEEE Access 9:98992–99008. https://doi.org/10.1109/ACCESS.2021.3095420

Khamassi I, Sayed-Mouchaweh M, Hammami M, Ghédira K (2018) Discussion and review on evolving data streams and concept drift adapting. Evol Syst 9:1–23. https://doi.org/10.1007/s12530-016-9168-2

Kleinberg J, Tardos E (1999) Approximation algorithms for classification problems with pairwise relationships: metric labeling and Markov random fields. Ann Symp Found Comput Sci Proc 49:14–23. https://doi.org/10.1109/sffcs.1999.814572

Kolter JZ, Maloof MA (2003) Dynamic weighted majority: A new ensemble method for tracking concept drift. In: Proceedings - IEEE international conference on data mining, ICDM. pp 123–130

Krawczyk B, Woźniak M (2015) One-class classifiers with incremental learning and forgetting for data streams with concept drift. Soft Comput 19:3387–3400. https://doi.org/10.1007/s00500-014-1492-5

Krawczyk B, Minku LL, Gama J et al (2017) Ensemble learning for data stream analysis: a survey. Inf Fus 37:132–156. https://doi.org/10.1016/j.inffus.2017.02.004

Li Z, Huang W, Xiong Y et al (2020) Incremental learning imbalanced data streams with concept drift: the dynamic updated ensemble algorithm. Knowledge-Based Systems 195:105694. https://doi.org/10.1016/j.knosys.2020.105694

Loo HR, Marsono MN (2016) Online network traffic classification with incremental learning. Evol Syst 7:129–143. https://doi.org/10.1007/s12530-016-9152-x

Lu J, Liu A, Dong F et al (2019) Learning under concept drift: a review. IEEE Trans Knowl Data Eng 31:2346–2363

Lughofer E, Angelov P (2011) Handling drifts and shifts in on-line data streams with evolving fuzzy systems. Applied Soft Computing 11:2057–2068. https://doi.org/10.1016/j.asoc.2010.07.003

Maciel BIF, Santos SGTC, Barros RSM (2015) A lightweight concept drift detection ensemble. https://doi.org/10.1109/ICTAI.2015.151

Mahdi OA, Pardede E, Ali N, Cao J (2020) Diversity measure as a new drift detection method in data streaming. Knowledge-Based Systems 191: 105227. https://doi.org/10.1016/j.knosys.2019.105227

Merigó JM, Pedrycz W, Weber R, de la Sotta C (2018) Fifty years of information sciences: a bibliometric overview. Inf Sci (n Y). https://doi.org/10.1016/j.ins.2017.11.054

Minku LL, White AP, Yao X (2010) The impact of diversity on online ensemble learning in the presence of concept drift. IEEE Trans Knowl Data Eng 22:730–742. https://doi.org/10.1109/TKDE.2009.156

Nordahl C, Boeva V, Grahn H, PerssonNetz M (2022) Evolvecluster: an evolutionary clustering algorithm for streaming data. Evol Syst 13:603–623. https://doi.org/10.1007/s12530-021-09408-y

Pesaranghader A, Viktor HL (2016) Fast Hoeffding drift detection method for evolving data streams. 96–111. https://doi.org/10.1007/978-3-319-46227-1

Pesaranghader A, Viktor HL, Paquet E (2018) McDiarmid drift detection methods for evolving data streams. In: Proceedings of the international joint conference on neural networks

Plamen A, Dimitar PF, Nik K (2010) Evolving Intelligent Systems: Methodology and Applications. Wiley-IEEE Press, United States.

Pratama M, Lu J, Lughofer E et al (2017) An incremental learning of concept drifts using evolving Type-2 recurrent fuzzy neural networks. IEEE Trans Fuzzy Syst 25:1175–1192. https://doi.org/10.1109/TFUZZ.2016.2599855

Pratama M, Pedrycz W, Lughofer E (2018) Evolving ensemble fuzzy classifier. IEEE Trans Fuzzy Syst 26:2552–2567. https://doi.org/10.1109/TFUZZ.2018.2796099

Qiao J, Sun Z, Meng X (2023) Interval type-2 fuzzy neural network based on active semi-supervised learning for non-stationary industrial processes. IEEE Trans Autom Sci Eng. https://doi.org/10.1109/TASE.2023.3237840

Ren S, Liao B, Zhu W, Li K (2018) Knowledge-maximized ensemble algorithm for different types of concept drift. Inf Sci (n Y) 430–431:261–281. https://doi.org/10.1016/j.ins.2017.11.046

Sakthithasan S, Pears R, Koh YS (2013) One pass concept change detection for data streams. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)

Santos SGTC, Barros RSM, Gonçalves PM (2019) A differential evolution based method for tuning concept drift detectors in data streams. Inf Sci (n Y) 485:376–393. https://doi.org/10.1016/j.ins.2019.02.031

Schlimmer JC, Granger RH (1986) Incremental learning from noisy data. Mach Learn. https://doi.org/10.1023/A:1022810614389

Sidhu P, Bhatia MPS (2019) A two ensemble system to handle concept drifting data streams: recurring dynamic weighted majority. Int J Mach Learn Cybern 10:563–578. https://doi.org/10.1007/s13042-017-0738-9

Souto R, de Barros M, Garrido S, Santos TDC (2019) An overview and comprehensive comparison of ensembles for concept drift. Inf Fus 52:213–244. https://doi.org/10.1016/j.inffus.2019.03.006

Street WN, Kim Y (2001) A streaming ensemble algorithm (SEA) for large-scale classification. Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '01 4:377–382. https://doi.org/10.1145/502512.502568

Suárez-Cetrulo AL, Quintana D, Cervantes A (2023) A survey on machine learning for recurring concept drifting data streams. Expert Systems with Applications 213:118934. https://doi.org/10.1016/j.eswa.2022.118934

Synnestvedt MB, Chen C, Holmes JH (2005) CiteSpace II: visualization and knowledge discovery in bibliographic databases. AMIA Annual Symposium proceedings 2005:724–728

van Eck NJ, Waltman L (2010) Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics. https://doi.org/10.1007/s11192-009-0146-3

Wang H, Xu Z, Zeng XJ (2018a) Modeling complex linguistic expressions in qualitative decision making: an overview. Knowl Based Syst. https://doi.org/10.1016/j.knosys.2017.12.030

Wang S, Minku LL, Yao X (2018b) A systematic study of online class imbalance learning with concept drift. IEEE Trans Neural Netw Learn Syst 29:4802–4821. https://doi.org/10.1109/TNNLS.2017.2771290

Wang X, Xu Z, Su SF, Zhou W (2021) A comprehensive bibliometric analysis of uncertain group decision making from 1980 to 2019. Inf Sci (n Y) 547:328–353. https://doi.org/10.1016/j.ins.2020.08.036

Wang S, MacHida F (2021) A robustness evaluation of concept drift detectors against unreliable data streams. 7th IEEE world forum on internet of things, WF-IoT 2021 569–574. Doi: https://doi.org/10.1109/WF-IoT51360.2021.9595202

Wares S, Isaacs J, Elyan E (2019) Data stream mining: methods and challenges for handling concept drift. SN Appl Sci 1:1–19. https://doi.org/10.1007/s42452-019-1433-0

White HD (2018) Pennants for garfield: bibliometrics and document retrieval. Scientometrics. https://doi.org/10.1007/s11192-017-2610-9

Widmer G, Kubat M (1996) Learning in the presence of concept drift and hidden contexts. Mach Learn. https://doi.org/10.1007/BF00116900

Yu D, Xu Z, Pedrycz W, Wang W (2017) Information sciences 1968–2016: a retrospective analysis with text mining and bibliometric. Inf Sci (n Y). https://doi.org/10.1016/j.ins.2017.08.031

Žliobaitė I (2010) Learning under concept drift: an overview. 1–36. https://doi.org/10.1002/sam