Artificial Intelligence Review
https://doi.org/10.1007/s10462-021-10087-z

Hierarchical classification of data streams: a systematic literature review

Eduardo Tieppo[1,2] · **Roger Robson dos Santos**[2] · **Jean Paul Barddal**[2] · **Júlio Cesar Nievola**[2]

Abstract

The classification task usually works with flat and batch learners, assuming problems as stationary and without relations between class labels. Nevertheless, several real-world problems do not assume these premises, i.e., data have labels organized hierarchically and are made available in streaming fashion, meaning that their behavior can drift over time. Existing studies on hierarchical classification do not consider data streams as input of their process, and thus, data is assumed as stationary and handled through batch learners. The same can be said about works on streaming data, as the hierarchical classification is overlooked. Studies concerning each area individually are promising, yet, do not tackle their intersection. This study analyzes the main characteristics of the state-of-the-art works on hierarchical classification for streaming data concerning five aspects: (i) problems tackled, (ii) datasets, (iii) algorithms, (iv) evaluation metrics, and (v) research gaps in the area. We performed a systematic literature review of primary studies and retrieved 3,722 papers, of which 42 were identified as relevant and used to answer the aforementioned research questions. We found that the problems handled by hierarchical classification of data streams include mainly classification of images, human activities, texts, and audio; the datasets are mostly created or synthetic data; the algorithms and evaluation metrics are well-known techniques or based on those; and research gaps are related to dynamic context, data complexity, and computational resources constraints. We also provide implications for future research and experiments to consider common characteristics shared amongst hierarchical classification and data stream classification.

Keywords Data stream mining · Hierarchical classification · Systematic literature review · Machine learning

We thank the Brazilian Research Support Agency Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) for their financial support.

✉ Eduardo Tieppo
eduardo.tieppo@ifpr.edu.br

[1] Instituto Federal do Paraná (IFPR) - Campus Pinhais, Pinhais, Brazil

[2] Pós-Graduação em Informática (PPGIa), Pontifícia Universidade Católica do Paraná (PUCPR), Curitiba, Brazil

Published online: 22 October 2021

1 Introduction

In Hierarchical Classification, machine learning models receive data examples associated with a path of labels representing classes and their hierarchical relationships (Silla and Freitas 2011; Freitas and Carvalho 2007; Wu et al. 2005). This hierarchical structure of classes is useful to represent several contexts, such as recognition of music genres and subgenres (Burred and Lerch 2003), computer-aided diagnosis where diseases are categorized by their etiology (Yassin et al. 2018), or recognition of animals organized in a taxonomy (Parmezan et al. 2018; Wu et al. 2005).

Understanding the hierarchical structure of a learning problem is helpful to split the information set into smaller sets and decisions (generic level classes), so these small decisions are made upon the relevant part of the data. This strategy allows a more fine-grained classification of a much larger number of classes that are fitted to the data since classifiers do not have to discriminate all classes together, but they use the class hierarchy information to lead the learning process (Dumais and Chen 2000; Silla and Freitas 2011).

Traditional hierarchical classification techniques are applied to static datasets and assume that all data will be available to a training step in the learning process. However, this assumption no longer reflects many of the real-world scenarios where data streams are produced continuously at a fast rate and impose time and memory constraints to learning models (Quiñonero-Candela et al. 2009; Nguyen et al. 2015).

Predictive models built based on static data may experience declines in performance as the data that continuously is made available may change its distribution, a phenomenon called concept drift (Widmer and Kubat 1996). Therefore, predictive models are required to deal with ever-changing environments and with real computational resource constraints since streaming can generate virtually infinite data input (Quiñonero-Candela et al. 2009; Nguyen et al. 2015).

The natural intersection between hierarchical classification and data streams results in several new challenges as the constraints of both contexts must be simultaneously considered. Moreover, current state-of-the-art techniques are not able to deal with these new problems directly. Hierarchical classification techniques cannot handle changing and potentially infinite data, while data stream classification techniques do not account for hierarchical relationships on classes from data samples.

Three comprehensive reviews on hierarchical classification were presented by Freitas and Carvalho (2007), Silla and Freitas (2011), and Defiyanti et al. (2019) in which grounding concepts and terminologies of the area were formally defined. Similarly, comprehensive reviews of data stream classification were presented in (Babcock et al. 2002), (Gaber et al. 2005) and (Gama 2010), showing that a fair amount of effort has been devoted to scenarios where data are made available as a stream and how its challenges, i.e., concept drift detection and adaptation as well as computational resource constraints, can be tackled.

These studies even show successful approaches and future challenges in both areas but do not present a perspective comprising both areas together and, principally, do not follow the rigor that a Systematic Literature Review (SLR) requires, as presented in (Wen et al. 2012).

In other words, despite research conducted on hierarchical classification and data stream classification separately, there is a lack of studies in hierarchical classification of data streams that consider the main characteristics of these kinds of problems together, such as hierarchical non-stationary data and context.

In this study, we present a Systematic Literature Review (SLR) focused on understanding the main characteristics of the hierarchical classification of data streams state-of-the-art. The purpose of this SLR is to summarize and clarify: (i) what kind of problems are handled by hierarchical classification of data streams, (ii) which datasets are frequently used in experiments in the existing studies, (iii) which algorithms, and (iv) evaluation metrics are used in the hierarchical classification of data streams and (v) what are the research gaps in the hierarchical classification of data streams.

Considering the previously described context, the main contribution of this study is three-fold:

- We present a systematic review comprising hierarchical classification of data streams, i.e., an intersection area between two well-known classification tasks: hierarchical classification and data stream classification. The review covers the main problems handled by hierarchical classification of data streams, datasets, classification algorithms and evaluation metrics used, and research gaps discussed in the studies.
- Hierarchical classification of data stream task has been addressed only superficially and sparingly in the literature. In this study, we provide formal definitions of this specific classification task.
- We provide implications for research and experiments and suggest directions for future works in hierarchical classification of data streams.

The remainder of this paper is organized as follows. Section 2 presents relevant definitions related to hierarchical classification, data stream classification and their intersection: the hierarchical data streams classification task. Section 3 describes the material and methods used in the systematic review. Sect. 4 contains the results of the systematic review and the discussions related to them, including the answers to the research questions. Section 5 concerns the implications for new research and experiments. Finally, Sect. 6 depicts the concluding remarks of this study.

2 Theoretical background

In this section, we provide definitions of both hierarchical and data stream classification tasks. Furthermore, we provide formal definitions of the intersection resulting task, i.e., the hierarchical classification of data streams.

2.1 Hierarchical classification

In hierarchical classification, instances are assigned to an independent label (class) that is part of a label path, where inner labels in the path represent hierarchical relationships with the outer labels. Classes are arranged in hierarchical structures, where nodes represent the classes and specialized nodes represent specific classes of its general nodes (Freitas and Carvalho 2007; Silla and Freitas 2011; Defiyanti et al. 2019).

Figure 1 compares a general approach of (a) flat classification and (b) hierarchical classification in an illustrative problem. In flat classification, the decision must be made while considering all the classes of the problem (all the possible song genres). Meanwhile, the hierarchical classification concerns an existent class taxonomy, which can be used to

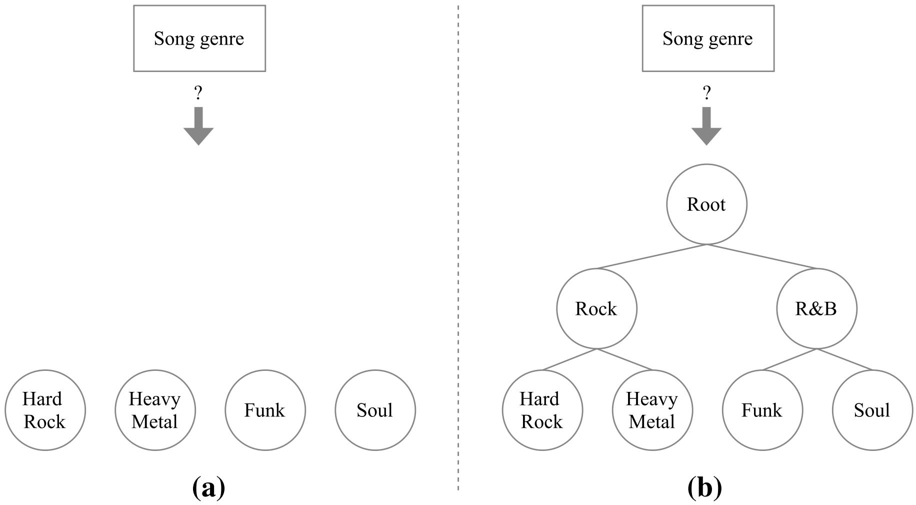

Fig. 1 Example of general approaches of **a** Flat Classification and **b** Hierarchical Classification in a hypothetical music genre problem. The class taxonomy can be used to lead to specific decisions about the classes by splitting the context complexity

make first smaller and generic decisions about the problem (in the example, to decide first between Rock and R&B genres), and then the specific ones.

A class taxonomy can be formalized as a regular concept hierarchy (Lu 1997) under a partially ordered set (Y,≻), where Y represents a finite set containing all target classes of a problem and the relation ≻ is defined as a subsumption relation ("is-a" relation) (Wu et al. 2005; Silla and Freitas 2011; Defiyanti et al. 2019).

According to Silla and Freitas (2011), a hierarchical classification problem can be categorized in a 3-tuple (Υ, Ψ, Φ), where:

- Υ specifies the data structure used to represent the class taxonomy and may be modeled using a Tree or a Directed Acyclic Graph (DAG) representation, according to how many parent nodes the same node has.
- Ψ defines the label cardinality, where instances of a given problem can have only one single path of labels (SPL) associated with them or multiple paths of labels (MPL); and,
- Φ describes the label depth, where problems support partial depth labeling (PD), or actual classes of the problem are represented only in the leaf nodes with full depth labeling (FD).

Similarly, a hierarchical classification algorithm can be categorized in a 4-tuple (Ω, Δ, Ξ, Θ), where:

- Ω specifies if the algorithm supports Tree or Directed Acyclic Graph (DAG) as data structures;
- Δ indicates if the algorithm can assign to an instance at most one predicted label path (single path prediction - SPP), or it can potentially assign multiple predicted label paths (multiple path prediction - MPP);

- $\varXi$ specifies if the algorithm always assigns leaf node classes as the last class of a predicted label path (mandatory leaf-node prediction - MLNP) or if it can predict label paths where the deeper class is at any hierarchy level (non-mandatory leaf-node prediction - NMLNP);
- $\varTheta$ describes how the hierarchical classifier handles the class hierarchy in its algorithm, comprising Local classifier per node (LCN), Local classifier per level (LCL), Local classifier per parent node (LCPN), or Global classifier (GC). In the LCN approach, one binary classifier per class handles each class in the hierarchy (except the root node). In the LCPN approach, one multi-class classifier per class (except on the leaf nodes) predicts between its child nodes. In the LCL approach, one multi-class classifier per level predicts between all nodes at the same level. Finally, in the GC approach, one single multi-class classifier is built to handle all classes using the hierarchy information.

Note that both categorizations are similar but use two different contexts: problems and algorithms. Nevertheless, some algorithm categories may be more suited to deal with specific categories of problems. For instance, a problem with full depth labeling (FD) should use a Mandatory leaf-node prediction scheme (MLNP) to be able to predict the entire path of labels since predicting only part of it would not represent an actual class of the problem (Silla and Freitas 2011).

The aforementioned categorization and definitions are still used nowadays, as seen in more recent works such as in Venkatesan and Er (2014), Kosmopoulos et al. (2015), Melo et al. (2017), Defiyanti et al. (2019), and Pereira et al. (2021).

Finally, Kiritchenko and Famili (2005) proposed three metrics able to measure the performance of a hierarchical classifier: hierarchical precision (*hP*), hierarchical recall (*hR*) and hierarchical F-Measure (*hF*). These metrics are variations of the traditional classification metrics (Precision, Recall and F-Measure) but instances are associated with a path of labels and the entire path is evaluated.

The *hP* metric, depicted in Eq. 1, computes the number of labels in a predicted label path ($\widehat{y}_i$) that are also components of the ground-truth label path (y_i) for the *i*-th instance. On the other hand, *hR*, depicted in Eq. 2, quantifies the number of ground-truth labels comprised by the predicted label path for a given instance.

$$hP = \frac{\sum_i |\widehat{y}_i \cap y_i|}{\sum_i |\widehat{y}_i|} \tag{1}$$

$$hR = \frac{\sum_i |\widehat{y}_i \cap y_i|}{\sum_i |y_i|} \tag{2}$$

Similarly to traditional classification metrics, the hierarchical F-Measure, depicted in Eq. 3, is the harmonic mean between hierarchical precision (*hP*) and hierarchical recall (*hR*). As in the traditional F-Measure, β weights *hP* and *hR* values (Cerri et al. 2015).

$$hF = (1 + \beta^2) \times \frac{hP \times hR}{(\beta^2 \times hP) + hR} \tag{3}$$

The interested reader can find a comprehensive review and seminal papers of hierarchical classification in (Silla and Freitas 2011) and (Kosmopoulos et al. 2015).

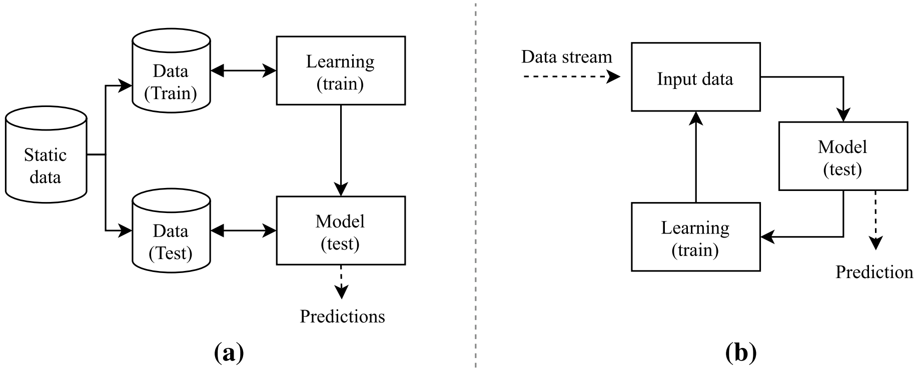

Fig. 2 Example of general approaches of **a** Traditional (batch) classification and **b** Data Stream Classification. An input data is obtained from the data stream, tested, incorporated into the model, and discarded; then, the cycle starts again

2.2 Data stream classification

Unlike traditional machine learning, where the dataset is static and can be accessed multiple times, data stream algorithms need to consider previously unmapped issues such as limited memory, single-pass data, readiness, and detection and adaptation to changes in the dataset (Quiñonero-Candela et al. 2009; Bifet and Kirkby 2009; Gama 2010; Gomes et al. 2019).

Figure 2 compares (a) a traditional classification process and (b) a data stream classification process. In traditional (or batch) classification, data are assumed to be static and completely available to the model at the training step; the dataset is then divided into subsets of training and testing data; the training data are submitted to the learning model that reviews them as many times as necessary, until obtaining a unique satisfactory model. This final model is then applied to the subset of testing data and provides predictions.

In contrast, in data stream classification, data are made available sequentially over time, and even a single instance can be provided to the model. This instance is tested by the model resulting in a prediction, and, only after that, it is incorporated into the model (being used as training data). Next, the cycle restarts with a new instance from the data stream. Any processed instance must be eventually discarded to maintain the model's ability to process new instances since the data stream is potentially unbounded.

According to the works of Gama (2010), Nguyen et al. (2015), and Widmer and Kubat (1996), due to the temporal and unbounded traits of data streams, learning algorithms must meet several constraints:

- Single-pass: each instance in the stream should be examined just once and cannot be reused;
- Readiness: learning and prediction should be made in real-time or near real-time;
- Bounded Memory: the amount of input data is gigantic or potentially infinite; therefore, a summary of the data stream is usually calculated and stored, and approximate results are acceptable;

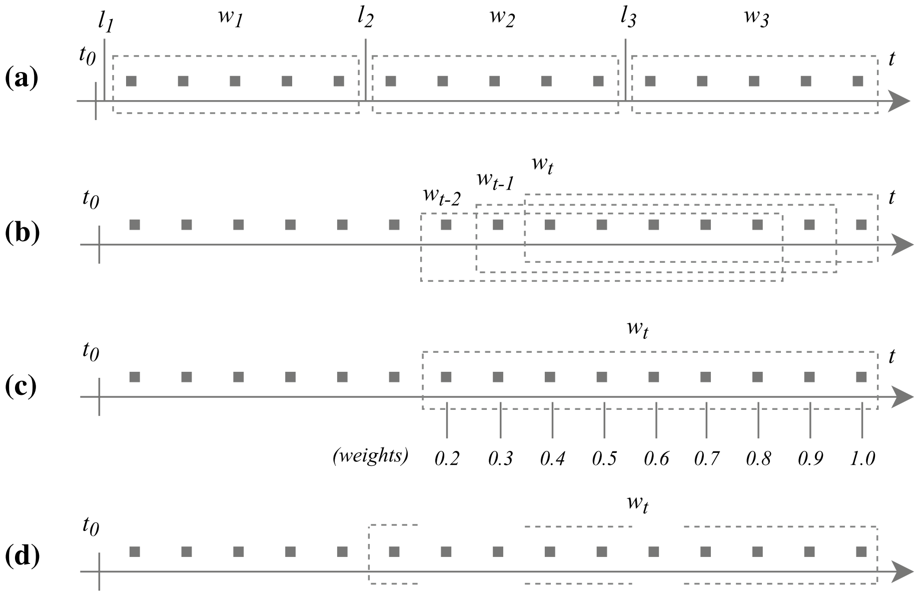

Fig. 3 Illustrations of different kinds of time windows: **a** Landmark window, **b** Sliding window, **c** Fading window and **d** Tilted-time window

- Concept drift detection: in data streams, concept drifts refer to the situation in which discovered patterns change over time; classification models should be able to detect and adapt to these drifts without human interference.

Since data streams are potentially unbounded, it is not possible to process all data at once. Thus, the data must be processed incrementally as data samples are made available. Data processing is performed according to different time window models, all of them sharing the same idea of heeding to specific portions of data (usually the most recent ones). These time windows can be of different types according to the way they process the data.

Figure 3 shows different time windows types. In the Landmark window (a), there is interest in portions of data between instances called “landmarks”; landmarks can be defined based on time, on the number of instances (mini-batches) and memory constraints; older or newer data have the same importance. In the Sliding window (b), there is more interest in the newer data and the time window slides along with time; data outside the window is discarded. In the Fading window (c), the data is weighted and there is more interest in newer data by assigning greater weights according to the data currency; thus, old data becomes less important in the learning process. Finally, in the Tilted-time window (d), there is interest in the most recent data by selecting instances based on elapsed time; the data are represented with different detail levels according to their age; thus, newer data has more instances and details and older data is represented with some smaller-scale pattern. It is also important to notice that the kinds of time windows are not limited to the ones illustrated in Fig. 3 and different kinds can even be mixed depending on the characteristics of the data (Nguyen et al. 2015; Barddal et al. 2017).

The main challenge in selecting a time window strategy is choosing a well-suited size due to the stability-plasticity dilemma (Mermillod et al. 2013). Shorter windows may help the model to become more responsive to drifts (plasticity). On the other hand, larger windows may result in more stable models (stability) (Gama 2010; Barddal et al. 2017).

Traditionally in data mining, algorithms work in batch (offline) mode. With large-scale data, this becomes infeasible due to memory and time limitations. To deal with that, algorithms in data stream classification can work with adaptations of batch mode (mini-batches), be incremental or adaptive (Gama et al. 2014).

Incremental algorithms regard updating or retraining the models using part of or all data as new instances becomes available. The Hoeffding Trees and Bayes models are examples of incremental algorithms (Gama et al. 2014; Domingos and Hulten 2000).

As a consequence of updating or retraining the models with eventually larger datasets, computational resources need to be tracked and evaluated. Nevertheless, as previously mentioned, concept drift also needs to be considered in data streams scenarios. In this case, models also need to adapt themselves to react to these drifts (Gama et al. 2014).

Adaptive algorithms also regard updating or retraining the models (as incremental ones), but the model includes strategies to forget the information previously learned. The adaptive models can be understood as advanced incremental learning models that are able to adapt to changes in data over time (Gama et al. 2014).

Concerning evaluation, authors in (Gama et al. 2014) proposed the prequential assessing method (or Interleaved Test-Then-Train (Bifet and Kirkby 2009)) to evaluate learning algorithms in streaming scenarios. In this process, each instance is used to test the model and the evaluation metrics, e.g., precision, recall, F-measure, are updated. Next, the instance is used to train/update the model. As metrics are calculated for each instance, results are often summarized using some strategy, such as using maximum or mean values and also considering some sampling frequency of the data stream.

Finally, there are still open research topics on the Data Stream Classification area. As models work in dynamic environments, they need to monitor themselves and respond accordingly. For instance, in the streaming context, models must detect and respond to concept drifts, monitor parameters and consider changes in the optimal setup, and even be prepared to deal with the appearance and disappearance of new features and classes over time, traits named feature and concept evolution (Gomes et al. 2019; Barddal et al. 2017).

In Concept Evolution, models must be able to detect the formation of a novel concept and consider it in an updated version of the model, including the redesign of the learning and evaluation steps. Similarly, the appearance of a new feature must be considered by data streaming methods, checking whether it is relevant or not and eventually incorporating it into the learning process. In both cases, this dynamic environment represents a key factor to the area since the automatic update of the computational representation of the learning model is challenging Masud et al. (2010); Gomes et al. (2019).

For more information on data stream mining, the interested reader can find comprehensive reviews and seminal papers in (Gama et al. 2013; Krempl et al. 2014; Nguyen et al. 2015; Ramírez-Gallego et al. 2017; Cano 2018).

2.3 Hierarchical classification of data streams

The Hierarchical Classification of Data Streams lies at the intersection of hierarchical classification and data streaming classification, two well-established research areas. Consequently, this new area inherits characteristics and challenges from its base areas and differs from traditional classification in two key aspects.

First, concerning hierarchical classification, examples must be assigned to not one independent label (class) but to a label path representing one of many possible label paths composing the class taxonomy. Next, concerning data stream classification, the entire dataset containing examples for a training step is not available; instead, examples are provided to the model sequentially over time.

Hierarchical classification of data stream methods must use data streams as input to their learning processes, not only as a source of data but effectively processing portions of the data over time using the assumption that there is no complete dataset.

Regarding input data, we denote $hDS = [(\mathbf{x}^t, \mathbf{y}^t)]_{t=0}^{\infty}$ to be a hierarchical data stream providing instances $(\mathbf{x}^t, \mathbf{y}^t)$, each of which arriving at a timestamp t, where $\mathbf{x}^t$ is a d-dimensional features set and its values, and $\mathbf{y}^t$ is the corresponding ground-truth label path (hierarchically structured classes) for the given instance. As mentioned above, class labels are organized under a regular concept hierarchy under a partially ordered set $(Y, \succ)$, where Y represents a finite set of all concepts and the relation $\succ$ is defined as an asymmetric, anti-reflexive and transitive subsumption ("is-a") relation (Silla and Freitas 2011).

The hierarchical classification of data streams can be formalized as a mapping function $f^t : \mathbf{x}^t \mapsto \mathbf{y}^t$, where a hypothesis f^t is continuously updated by features $\mathbf{x}$ to the corresponding labels y^t accurately (Gama 2010). Also, data streams are expected to be ephemeral due to the intrinsic time component, and, thus, the underlying data distribution is expected to change, a phenomenon named concept drift (Tsymbal 2004; Gama et al. 2014).

A concept (C) is defined as set of prior probabilities of the classes and class-conditional probability density function given by $C = \bigcup_{y \in Y} \{(P[y], P[\mathbf{x}|y])\}$ (Barddal et al. 2016; Nguyen et al. 2012). A concept drift occurs if, at between two timestamps t_i and $t_j = t_i + \Delta$ with $\Delta > 1$, $C^{t_i} \neq C^{t_j}$ holds (Tsymbal 2004; Barddal et al. 2016). Consequently, f^t should capture the data dynamics accordingly.

Finally, methods need to exhibit bounded computational resources (time and memory), examining each example only once according to their arrival and processing it in less time than the ratio in which new instances become available. Otherwise, the method will eventually need to drop incoming examples, or it will not be able to adapt itself quickly enough to handle concept drifts (Barddal et al. 2016; Bifet and Kirkby 2009).

3 Systematic literature review method

3.1 Protocol

Our review was based on the systematic literature review (SLR) method proposed by Kitchenham and Charters (2007) with minor adaptations applied in a machine learning systematic literature review proposed by Wen et al. (2012). The protocol used in this study includes seven steps: research questions definition, search strategy design, exclusion criteria, inclusion criteria, data extraction, quality assessment criteria, and data synthesis. These steps are layered in Fig. 4.

First, we defined the research questions based on the objectives of our SLR: summarize and clarify the main problems, datasets, algorithms, evaluation metrics, and research gaps in the hierarchical classification of data streams. In the second step, we designed a search strategy to select papers related to our research questions, thus defining a search query and target databases. This step resulted in multiple candidate studies retrieved from the query search in each of the databases. Next, these candidate studies were filtered according to inclusion and

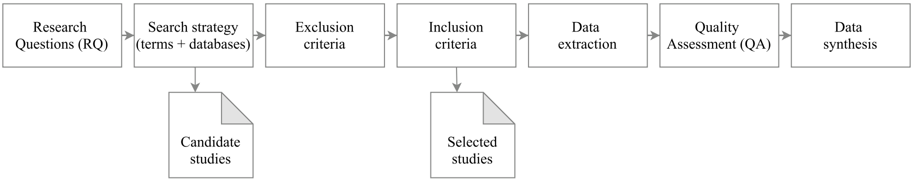

Fig. 4 Main steps of the systematic literature review proposed method. Adapted from Wen et al. (2012)

exclusion criteria (steps 3 and 4), resulting in a final subset of selected studies. In Step 5, the data from these selected studies was gathered using an extraction card and evaluated using Quality Assessment (QA) criteria in step 6. Finally, all the information collected was synthesized and analyzed (step 7). Each of the aforementioned steps is detailed as follows.

3.2 Research questions

To fulfill the objectives of this research, five Research Questions (RQ) were defined as follows:

- *RQ1*: Which problems are handled by hierarchical classification of data streams? Our goal is to list which kinds of problems are addressed by the studies in a hierarchical classification of data streams, e.g., audio processing, natural language processing, image recognition. The result of RQ1 is a list of categories of problems dealt with.
- *RQ2*: Which datasets are frequently used in the studies?
  RQ2 identifies which datasets (or types of datasets) are used in studies at the intersection of hierarchical classification and data streams. The result of RQ2 is a list of dataset names and sources used.
- *RQ3*: Which algorithms are used in experiments in the studies?
  RQ3 focuses on the identification of learning algorithms algorithms used or developed in studies concerning the hierarchical classification of data streams. RQ3 also comprises algorithms related to the classification tasks. The result of this RQ is a list of algorithms used in the studies and their basis.
- *RQ4*: Which evaluation metrics are used in experiments in the studies?
  RQ4 aims at identifying the metrics that have been used as evaluation criteria in studies concerning the hierarchical classification of data streams. The result of this RQ is a list of evaluation metrics used in the studies and their basis.
- *RQ5*: What are the research gaps in the hierarchical classification of data streams context?
  RQ5 is concerned with identifying possible research gaps in the hierarchical classification of data streams reported by the authors of the studies in their rationale, discussion, limitations or conclusions. The result of this RQ represents possible research paths for new studies.

3.3 Search strategy

To create the search query, we selected terms that describe both “Hierarchical Classification” and “Data Stream”. Synonyms and related terms were also added forming the final query:

- “Hierarchical Classification” AND (“Data Stream” OR “Data-Streaming” OR “Stream” OR “Streaming” OR “Online” OR

```
‶Online Data″OR ‶Real-time″ OR ‶Real-time Data″) AND (‶Machine Learning″ OR‶Data Mining″ OR ‶Pattern Recognition″)
```

The related term “machine learning” and synonyms have been added to avoid misleading results caused by the term “stream”, which can be found in articles related to other areas, such as ecology articles (running waters, for example).

The search was performed in July 2019 on five electronic databases: ScienceDirect[1], Scopus[2], SpringerLink[3], IEEE Xplore[4] and ACM Digital Library[5].

The search query was matched against full-text primary studies published in journals and conferences in any period of time.

3.4 Exclusion and inclusion criteria

The studies retrieved from the databases previously described in Sect. 3.3 resulted in the candidate studies. Candidate studies were filtered according to Exclusion Criteria (`EC`) to select only primary studies potentially able to answer the proposed research questions. The Exclusion Criteria followed are presented below:

- `EC1`: It is a duplicate study;
- `EC2`: It is not published in English;
- `EC3`: It is grey literature;
- `EC4`: It is a reprint of an original study.

Thus, the candidate studies were also filtered by the Inclusion Criteria (`IC`) to select only relevant studies to those proposed research questions. The Inclusion Criteria are listed below.

- `IC1`: It discusses hierarchical data stream problems from the machine learning perspective;
- `IC2`: It uses public datasets to represent hierarchical data stream classification problems;
- `IC3`: It proposes or uses data stream hierarchical classification algorithms;
- `IC4`: It proposes or uses data stream hierarchical classification evaluation metrics.

By applying Exclusion and Inclusion Criteria, we obtained a filtered set of studies from the original candidate studies, named selected studies. The application of the inclusion criteria process was conducted independently by two different authors of this study, performing a double-checking inclusion. In case of disagreement, we included the study.

The agreement between the authors was measured using the Cohen Kappa statistic (Cohen 1960). The final set of selected studies was submitted to the last three steps: data extraction, quality assessment, and data synthesis.

[1] https://www.sciencedirect.com/search/advanced.

[2] https://www.scopus.com/search/.

[3] https://link.springer.com/search.

[4] https://dl.acm.org/advsearch.cfm.

[5] https://ieeexplore.ieee.org/search/advsearch.jsp?expression-builder.

Table 1 Data extraction card

Basic information
Study Identifier
Source (journal or conference identification)
Link
Document Type (Article or Conference Paper)
DOI
Year
Authors
Title
Keywords
Abstract
Research questions
RQ1: Which problems are handled by hierarchical classification of data streams?
RQ2: Which datasets are frequently used in the studies?
RQ3: Which algorithms are used in the experiments in the studies?
RQ4: Which evaluation metrics are used in the experiments in the studies?
RQ5: What are the research gaps in the hierarchical classification of data streams context?
Quality assessment
QA1: Are the objectives well defined?
QA2: Are the research questions well defined?
QA3: Is there a research gap that justifies the study?
QA4: Are the methods suited for the objectives?
QA5: Are the methods well defined?
QA6: Is there results' validation?
QA7: Are the limitations explicitly presented?

3.5 Data extraction

We performed the data extraction process using an extraction card addressing basic information of the selected studies and the research questions concerned in this review.

Table 1 shows the extraction card used; the form was proposed, tested in a pilot data extraction with a sample of the selected studies, and then refined. The two first authors of this study extracted data, culminating in a double-checking extraction. In case of a disagreement, the authors discussed and, and if no consensus was reached, the remainder of the authors were involved for tie-breaking. In the case of different information extracted by researchers, we merged both extractions to avoid relevant information loss.

3.6 Study quality assessment

The quality assessment (QA) of the selected studies was designed to improve the reliability of the review protocol by working as a measure of the rigorousness of the selected papers.

We elaborated the following set of questions regarding methodological criteria about the studies:

- QA1: Are the objectives well defined?
- QA2: Are the research questions well defined?
- QA3: Is there a research gap that justifies the study?
- QA4: Are the methods suited for the objectives?
- QA5: Are the methods well defined?
- QA6: Is there results' validation?
- QA7: Are the limitations explicitly presented?

We answered the QA questions using a Likert scale from 1 to 5, where higher ratings represent a more fitted study to the evaluated criterion. Ratings of 5 indicate studies where the criterion is clearly defined or fitted, while ratings of 1 indicate the absence of the evaluated criterion.

Studies were assessed with a score of 5 when they presented the subject evaluated by the QA clearly defined and described and on a suitable study section; for example, when the objectives of the study (subject evaluated by the QA) were explicit and already described in the introduction of the paper. Studies assessed with a score of 4 presented the subject evaluated by the QA well defined but implicit in a section. We assessed with a score of 3 the studies that presented the corresponding subject defined throughout the entire text. Studies with a score of 2 contained the subjects deductible but not defined. Finally, studies assessed with a score of 1 did not present the subject or similar.

As in the inclusion criteria, two authors performed the quality assessment (QA) independently. The agreement between the authors was measured using the Cohen Kappa statistic (Cohen 1960), the average disagreement between the authors and the euclidean distance between the ratings provided by the authors.

It is important to notice that the quality assessment step does not aim at providing a ranking of studies and does not represent any kind of ground truth about the studies. The criteria work as a guide for interpreting the results, findings, or inferences shown in this review.

3.7 Data synthesis

The data synthesis step summarized extracted data and information that would contribute to a better understanding of the research questions concerned in this review and that could gather evidence to answer them. The extracted data included both qualitative and quantitative data, depending on which research question is analyzed. We used meta-analysis to summarize the results of independent studies (Higgins and Green 2011).

The minor data synthesis strategies used in each research question are detailed as follows. Concerning RQ1 and RQ5, we used a meta-ethnographic approach to synthesizing qualitative using reciprocal translation (Noblit and Hare 1988; Atkins et al. 2008). In RQ1, for instance, studies regarding phoneme sequence recognition (Hamooni et al. 2016), radio-program classification (Kotsakis et al. 2012), and radar pulse classification (Kauppi et al. 2010) can all be understood as audio related tasks. Concerning RQ2, RQ3, and RQ4, we also used reciprocal translation but using direct counting of major categories of the evaluated subject. In RQ3, for instance, the Support Vector Machine (SVM) with more

Table 2 Number of retrieved studies

Retrieved set of studies	Number of studies
Candidate studies	3722
Candidate studies (after exclusion criteria)	2849
Selected studies	42

middle vectors used in (Wang et al. 2009), the SVM binary tree approach proposed in (Xie et al. 2011) and the adapted version of SVM to text streams presented in (Ren et al. 2014) could all be understood as variants of the traditional SVM.

4 Results and discussion

This section overviews selected studies, answers the five research questions previously described, and summarizes study quality assessment results. The next subsection shows a quantitative analysis of the main characteristics of the selected studies, followed by other subsections organized according to the research questions. Finally, the last subsection shows quality assessment results.

4.1 Overview of selected studies

The first set of retrieved studies was obtained in the search strategy step (see Sect. 3.3). Next, the studies were filtered using the exclusion and inclusion criteria, resulting in two more subsets. Table 2 shows the number of studies encompassed in each step.

As previously described, the agreement between the authors was measured using Cohen's Kappa statistic (Cohen 1960). In the selected studies, a Cohen's Kappa index of 27.98% was observed, which represents a fair agreement according to the Cohen's Kappa guidelines (Landis and Koch 1977).

An overview of the final selected studies is given in Table 3.

After the search strategy step, we identified 42 studies related to hierarchical classification of data streams. These studies were published in the 2008-2019 period. Figure 5 shows the distribution of studies per year of publication. It is important to highlight that the number of articles published in 2019 (4) represents studies published until the date of study retrieval; therefore, this number can be greater if the whole year is considered.

Among the selected studies, 12 (29%) of them were published in conference proceedings and 30 (71%) of them in journals. The publication venues and distribution of the selected studies are shown in Table 4. Notice that the "Other" row represents publication venues in which only one of the selected studies was published. This may occur due to studies published in publication venues focused on the context of the problem dealt with by the study rather than the computational method employed. For instance, Korda et al. (2015) published a study of identification of oculomotor behavior in Computers in biology and medicine Journal; also, Puerto-Souza et al. (2015) published a study of classification of colonoscopy videos in the International Workshop on Computer-Assisted and Robotic Endoscopy.

Table 3 Selected studies

Study ID	Publication type	Year	Authors and reference	Title
S1	Article	2008	del Campo-Ávila et al. (2008)	Improving the performance of an incremental algorithm driven by error margins
S2	Article	2009	Gu et al. (2009)	An Adaptive Hierarchical Model Based on Fusion of Ontology and Context
S3	Conference paper	2009	Wang et al. (2009)	Hierarchical Classification of Business Information on the Web Using Incremental Learning
S4	Conference paper	2009	Song et al. (2009)	Problem classification method to enhance the ITIL incident and problem
S5	Article	2010	Kauppi et al. (2010)	Hierarchical classification of dynamically varying radar pulse repetition interval modulation patterns
S6	Article	2010	Chou et al. (2010)	Integrating support vector machine and genetic algorithm to implement dynamic wafer quality prediction system
S7	Article	2010	Lughofer (2010)	On-line evolving image classifiers and their application to surface inspection
S8	Conference paper	2010	Martin et al. (2010)	Soft Concept Hierarchies to Summarise Data Streams and Highlight Anomalous Changes
S9	Article	2011	Xie et al. (2011)	Pitch-density-based features and an SVM binary tree approach for multi-class audio classification in broadcast news
S10	Article	2012	Shi et al. (2012)	A method for classifying packets into network flows based on GHSOM
S11	Article	2012	Jeong and Lee (2012)	Adaptive object recognition model using incremental feature representation and hierarchical classification
S12	Conference paper	2012	Djorgovski et al. (2012)	Flashes in a star stream: Automated classification of astronomical transient events
S13	Article	2012	Song et al. (2011)	Hierarchical online problem classification for IT support services
S14	Article	2012	Kotsakis et al. (2012)	Investigation of broadcast-audio semantic analysis scenarios employing radio-programme-adaptive pattern classification
S15	Article	2014	La et al. (2014)	Classifying XML data of semantic sensor networks
S16	Article	2014	D'hondt et al. (2014)	Dealing with temporal variation in patent categorization
S17	Conference paper	2014	Ren et al. (2014)	Hierarchical multi-label classification of social text streams
S18	Article	2014	Purohit et al. (2014)	Identifying seekers and suppliers in social media communities to support crisis coordination
S19	Article	2015	Alazrai et al. (2015)	Anatomical-plane-based representation for human–human interactions analysis
S20	Article	2015	Korda et al. (2015)	Automatic identification of oculomotor behavior using pattern recognition techniques
S21	Article	2015	Fan et al. (2015)	Cost-sensitive learning of hierarchical tree classifiers for large-scale image classification and novel category detection
S22	Article	2015	Kasaei et al. (2015)	Interactive open-ended learning for 3d object recognition: An approach and experiments
S23	Conference paper	2015	Peng et al. (2015)	Trending sentiment-topic detection on twitter

Table 3 (continued)

Study ID	Publication type	Year	Authors and reference	Title
S24	Conference paper	2016	Peixoto et al. (2016)	Adaptive learning process for the evolution of ontology-described classification model in big data context
S25	Conference Paper	2016	Chavez et al. (2016)	Automated species counting using a hierarchical classification approach with Haar cascades and multi-descriptor random forests
S26	Conference paper	2016	Puerto-Souza et al. (2015)	Enhancing normal-abnormal classification accuracy in colonoscopy videos via temporal consistency
S27	Article	2016	Chen et al. (2016)	Human continuous activity recognition based on energy-efficient schemes considering cloud security technology
S28	Article	2016	Weigl et al. (2016)	On improving performance of surface inspection systems by online active learning and flexible classifier updates
S29	Article	2016	Hamooni et al. (2016)	Phoneme sequence recognition via DTW-based classification
S30	Conference paper	2017	Anderez et al. (2017)	A hierarchical approach towards activity recognition
S31	Article	2017	Chakroun et al. (2017)	SW-SGD: the sliding window stochastic gradient descent algorithm
S32	Conference paper	2018	Silva-Palacios et al. (2018)	Adapting Hierarchical Multiclass Classification to changes in the target concept
S33	Article	2018	Sun et al. (2018)	Affect recognition from facial movements and body gestures by hierarchical deep spatio-temporal features and fusion strategy
S34	Article	2018	Khowaja et al. (2018)	Contextual activity based Healthcare Internet of Things, Services, and People (HIoTSP): An architectural framework for healthcare monitoring using wearable sensors
S35	Article	2018	Liu et al. (2018)	Fine-grained entity type classification with adaptive context
S36	Article	2018	Cao et al. (2018)	GCHAR: An efficient Group-based Context-aware human activity recognition on smartphone
S37	Article	2018	Pesaranghader et al. (2018)	Reservoir of diverse adaptive learners and stacking fast hoeffding drift detection methods for evolving data streams
S38	Article	2018	Protasov et al. (2018)	Using deep features for video scene detection and annotation
S39	Article	2019	Huang et al. (2018)	A crowdsource-based sensing system for monitoring fine-grained air quality in urban environments
S40	Article	2019	Raza et al. (2019)	Diagnosis and Monitoring of Alzheimer's Patients Using Classical and Deep Learning Techniques
S41	Article	2019	Saggese et al. (2019)	Learning skeleton representations for human action recognition
S42	Conference paper	2019	Parmezan et al. (2018)	Towards Hierarchical Classification of Data Streams

Regarding the topics addressed by the studies, Figs. 6 and 7 show the top ten (including ties) keywords of the selected studies. Figure 6 shows the most used words into the keywords set, while Fig. 7 shows the actual keywords. The most common keywords are “Machine Learning”, “Incremental Learning”, and “Hierarchical classification” (7, 6, and 6 times, respectively). It is important to notice that, in this context, words like “Twitter” or “Online” represent data stream sources/environments.

4.2 Problems handled by hierarchical classification of data streams studies (RQ1)

The main problems addressed in these studies are related to the classification of images, human activities, text, and audio. Figure 8 shows these categories and other ones less addressed. Some studies handled more than one kind of problem and others did not focus on any specific area. Most studies did not base their methods on the type of problem addressed, except for a few studies in which the classification techniques themselves were domain-dependent, such as in (Liu et al. 2018), where the authors have used natural language processing techniques, which are not directly transposable to other domains than textual without major changes.

4.3 Most used datasets (RQ2)

We observed a multitude of datasets used in the hierarchical classification of data streams studies. Table 5 shows the categorization of the datasets used in the studies regarding the source of the data. We used the term “Obtained from Literature” to refer to studies that used datasets previously published in other studies or available in public repositories. The term “created to the study” represents datasets that were made by the authors to be used as input data in the experiments in the studies. Finally, the term “Synthetic” represents datasets also produced by the authors but using simulation techniques to produce the data.

It is important to clarify that Table 5 only shows the kinds of datasets used in the studies and not the total number of datasets. The reader can find an exhaustive list of the datasets used in each study in Table 10 given in "Appendix (A)".

4.4 Most used algorithms (RQ3)

The algorithms used in the selected studies are shown in Fig. 9. As described above, the figure summarizes the algorithms used by the studies showing the technique used as the basis for the algorithm application.

Algorithms based on other kinds of techniques are counted on the “Other” row. As is RQ2, the reader can find an exhaustive list of the algorithms related to each study in Table 10 (A).

We observe that the main techniques used at the core of the proposed methods are based on traditional machine learning techniques, even though they have to deal with a new kind of problem with specific characteristics. The traditional k-Nearest Neighbors (kNN) technique (Aha et al. 1991), for instance, is not well suited for the streaming scenario since it has to make comparisons between large quantities of data to determine the nearest neighbors, which is potentially forbidden in computational terms when dealing with virtually infinite data. We can also observe that tree-based algorithms (Decision Tree (Quinlan

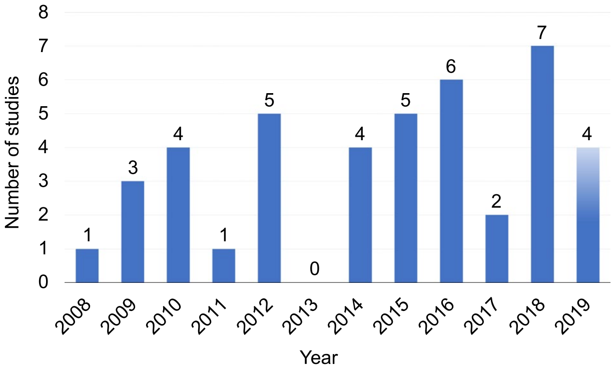

Fig. 5 Distribution of the papers per year

Table 4 Publication venues and distribution of selected studies

Publication venue	Number of studies
Neural networks	3
Pattern recognition	2
Expert systems with applications	2
Other (only 1 study per publication venue)	35

2014) and Random Forests (Breiman 2001)) appear in the list but without usual algorithms in the streaming scenario, such as the Hoeffdding Trees (Domingos and Hulten 2000).

4.5 Most used evaluation metrics (RQ4)

The evaluation metrics used in the selected studies are shown in Fig. 10. Despite the need for most representative metrics for Hierarchical Classification, e.g., hP, hR and hF; and for Data Stream classification, e.g., memory consumption; we can observe that the plain formulations of Accuracy, Precision, Recall and F-Score metrics are still the most used metrics by the selected studies, plus the loss and time consumption metrics.

Other evaluation metrics used only once in the studies are counted on the "Other" item. The reader can also find an exhaustive list of the evaluation metrics used in each study in Table 10 ("Appendix A").

4.6 Research gaps in the hierarchical classification of data streams context (RQ5)

As previously described, our review addressed RQ5 through meta-ethnographic and reciprocal translation. We were concerned about looking for problems handled by authors during experiments, limitations that provided clues about major context problems, and future works explicitly reported by the authors that represented possible research gaps. We found that the research gaps are related to the following three main topics: (i) dynamism, (ii) data complexity, and (iii) computational resources.

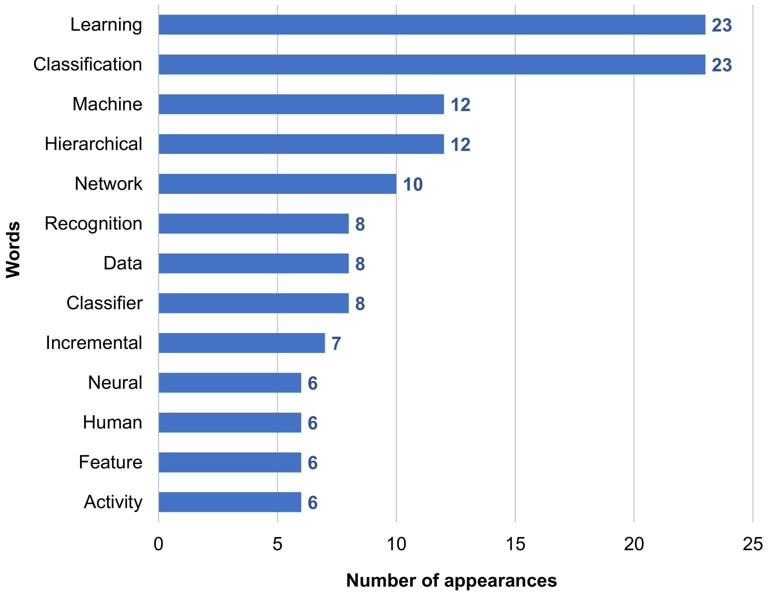

Fig. 6 Main terms used as keywords of the selected studies

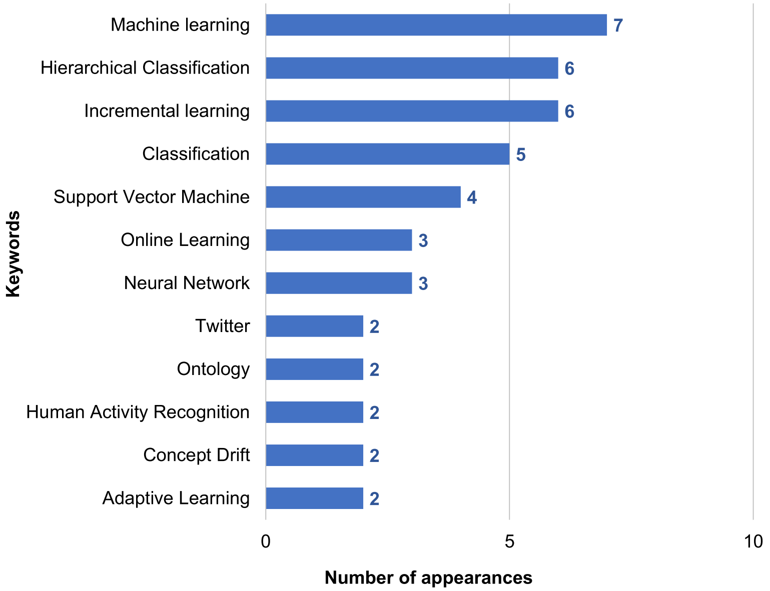

Fig. 7 Main keywords used in the selected studies

First, several studies have reported the dynamic environment when handling hierarchical classification of data streams as an additional difficulty at every step of the process.

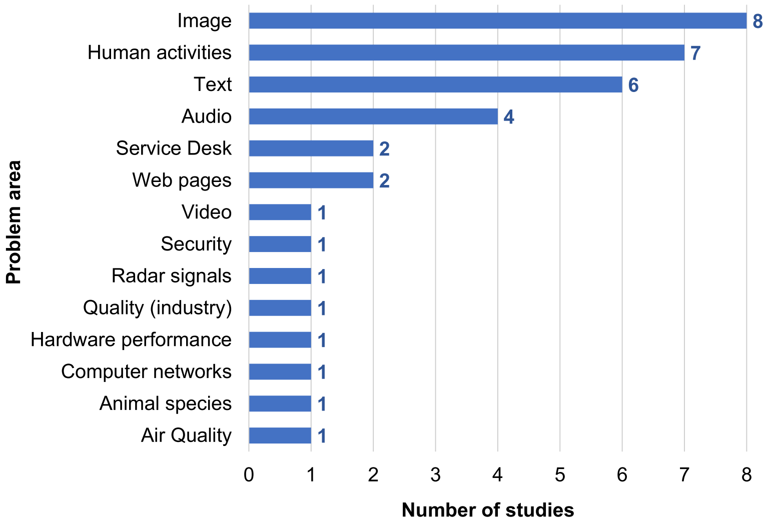

Fig. 8 Categories of problems handled by the studies

Table 5 Types of datasets used in the studies

Dataset type	Number of studies
Obtained from literature	17
Created to the study	19
Synthetic	4

In the work of (Wang et al. 2009), for instance, the authors proposed an adapted incremental SVM to classify web pages, but their model loses representativeness when new data is used for model updating, as it considers only the new data without considering support vectors learned previously.

Another example related to model adaptation can be found on the comprehensive analysis about concept drift in patent texts presented in (D'hondt et al. 2014), where authors report that one of the biggest difficulties for classification models that work with data over time is deciding how to incorporate unseen/new data into the classification models.

As another example of the dynamism in the models, the authors in (Weigl et al. 2016) proposed an active learning method to classify surface images to respond more accurately in changing production environments, but they cite the difficulty in handling concept drift on the fly as a major limitation of their method.

In the studies presented in (Kotsakis et al. 2012) and (Peng et al. 2015) (audio semantic analysis and text sentiment analysis, respectively), the authors also related a need for methods that are able to dynamically extract and select features based on the context of the problem or over time.

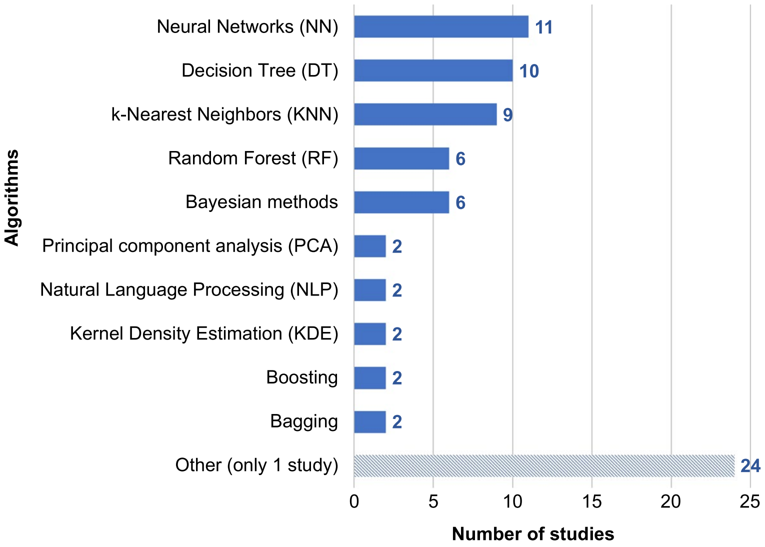

Fig. 9 Algorithms used in the studies

In the work (Silva-Palacios et al. 2018), the authors handled changes in concept by adapting their proposed model adding new classes but reported a consequent side effect: data imbalance in new classes.

The second major topic related to research gaps regards data complexity.

First, hierarchically structured data bring other levels of complexity to problems, such as related data classes, a possible multiple path prediction or non-mandatory leaf node prediction, imbalanced data, etc (Silla and Freitas 2011).

Second, data stream as input culminates in a massive number of instances, not always labeled, and evolving over time (Barddal et al. 2017). In the work of (Jeong and Lee 2012), the authors reported the difficulty in generating generic knowledge without having access to complete data as one of the main barriers to obtaining better results. Similarly, the authors in (Djorgovski et al. 2012) cite that data are often sparse, heterogeneous, and incomplete, evolving in time.

Finally, the third obstacle in the hierarchical classification of data streams concerns computational resources. In (Wang et al. 2009), the authors mention as a limitation of their proposal the trade-off between higher accuracy rates and training time and memory consumption rates, which may be prohibitive in real streaming scenarios.

The authors in (Song et al. 2011) also establish a mandatory need for massive data to achieve higher prediction rates, yet, no ablation study has been conducted to justify such a claim.

In (Pesaranghader et al. 2018), the authors suggest that, in some cases, the proposed methods become impracticable in real conditions, and cite as an example works that use smartphone sensors as data sources but disregard the limited resources of this kind of device in their experiments.

4.7 Quality assessment overview

Table 6 shows the quality scores of the selected studies.

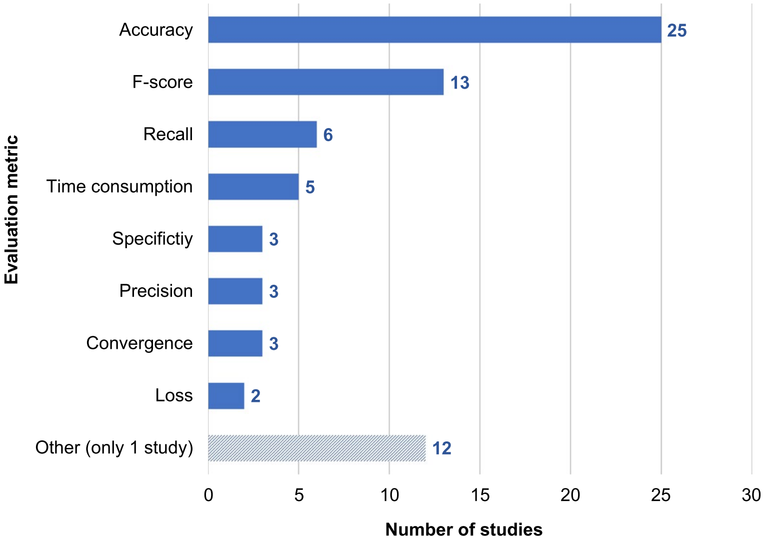

Fig. 10 Evaluation metrics used in the studies

As aforementioned, the two first authors of this study performed the quality assessment (QA) independently and are identified in the table as researchers "A" and "B". The second to last column shows the study score by researcher (an average of the individual scores of quality assessment questions). The last column depicts the final score (an average between the researchers) of each selected study.

Table 7 shows the average scores of researchers by quality assessment questions to elucidate the level of agreement between the researchers.

Similarly, Table 8 shows agreement between the researchers using the Cohen Kappa statistic, the average disagreement between them, and the euclidean distance between the scores provided by the researchers.

We emphasize that we measured the agreement between the authors to increase the reliability of the results and findings of our study by minimizing threats and biases in the evaluation using the same criteria for all the assessed studies.

Furthermore, as highlighted in Sect. 3.6, the quality scores should not be used to rank the studies. The scores serve as a guide for interpreting the results, findings, or inferences shown in this review. For instance, an interested reader will easily find answers to the research questions addressed by our review in those studies with higher assessment scores.

5 Discussion and State of the art

In this section, we discuss the adherence of selected studies to the Hierarchical Classification of Data Streams area. To that, we summarize the main research topics addressed or not covered by the studies; also, we present implications for further research and experiments in the area by discussing the main findings and research gaps indicated by the authors of the primary studies that compose this review.

5.1 Adherence of the selected studies

We categorized the selected studies according to the properties of Hierarchical and Data Stream Classification areas (see Sect. 2).

Table 9 details the adherence of the selected studies to the Hierarchical Classification of Data Stream properties. Columns 2-5 show properties inherited from the Hierarchical Classification area and columns 6-11 from the Data Stream Classification area. Unfilled cells (-) represent topics not comprehensively addressed by that study.

One can note that, even in the selected studies, there are few cases in which all the properties from the Hierarchical Classification of Data Streams area are fulfilled together.

Some studies present full hierarchical classifiers but only partially address the data streams aspect when considering data from streams as input while ignoring some constraints brought together by this data (like concept drifts) or only producing theoretical essays of their methods. On the other hand, some data stream classifiers even work with hierarchical data streams but perform their classification process by ignoring the hierarchy and obtaining a flat representation of classes.

As an exception, the study presented in (Parmezan et al. 2018) was able to cover both areas of Data Stream Classification and Hierarchical Classification concomitantly by proposing a method at the intersection of the areas from the beginning and also making available three datasets of Hierarchical Classification of Data Streams.

Their method is based on the k-Nearest Neighbors (kNN) technique, represents the data hierarchically, and classifies new data using a top-down strategy within the hierarchy. The proposed algorithm performs a single path (Δ) and non-mandatory leaf-node (Ξ) predictions, represents the hierarchy in a tree (Ω), and uses a local classifier per parent node approach to handle the hierarchy (Θ). In addition, the algorithm uses a sliding window by storing a memory buffer on each class node with the most recent examples of the data stream and discarding older instances when the buffer is full. After a predetermined number of initial instances used for training (burnout window), the method follows the prequential assessing protocol, processing the data stream on an instance basis and discarding each instance after analyzing it.

By this approach, the method successfully merged both areas (data stream classification and hierarchical classification) being able to classify hierarchical data streams with bounded computational resources and responsive to possible concept drifts. However, the method still has some limitations in the context of Hierarchical Classification of Data Streams since the computational cost for classifying new instances is dependent on the number of instances that the model stores.

Consequently, the available computational resources still limit the method in representing information since larger buffers could compromise its readiness and even make it unfeasible because the processing time will eventually surpass the ratio in which incoming instances become available. Then, the method will need to discard those instances without processing them, and it will not adapt itself swiftly enough to handle concept drifts since information about changes in the data would be lost.

In this sense, there is room for methods that fulfill the requirements of the Hierarchical Classification of Data Streams area and can work with constant computational resources and not only with resources linearly scalable by the number of instances provided by the data stream.

Table 6 Quality scores of selected studies

Study ID	Researcher	QA 1	QA 2	QA 3	QA 4	QA 5	QA 6	QA 7	Score by researcher	Final score
	A	3	2	2	2	3	3	2	2.4	
S1	B	3	1	2	2	2	3	2	2.1	2.3
	A	–	–	–	–	–	–	–	–	
S2	B	–	–	–	–	–	–	–	–	–
	A	3	2	4	1	3	2	3	2.6	
S3	B	3	1	5	2	3	3	2	2.7	2.6
	A	3	2	2	2	3	2	2	2.3	
S4	B	3	1	1	2	3	1	1	1.7	2.0
	A	3	2	2	3	3	3	1	2.4	
S5	B	3	1	2	3	3	4	2	2.6	2.5
	A	3	1	3	3	3	2	1	2.3	
S6	B	2	1	2	3	4	1	1	2.0	2.1
	A	2	1	3	3	4	3	3	2.7	
S7	B	3	1	2	3	5	4	3	3.0	2.9
	A	2	2	2	2	3	2	1	2.0	
S8	B	2	1	1	1	3	1	1	1.4	1.7
	A	3	2	1	2	3	3	2	2.3	
S9	B	2	1	1	2	4	3	1	2.0	2.1
	A	3	2	3	3	3	2	3	2.7	
S10	B	2	1	4	3	3	1	3	2.4	2.6
	A	2	1	3	4	4	3	2	2.7	
S11	B	3	1	4	4	5	3	1	3.0	2.9
	A	3	2	2	3	3	2	2	2.4	
S12	B	2	1	4	2	2	3	3	2.4	2.4
	A	3	2	2	3	2	2	2	2.3	
S13	B	3	1	1	2	3	2	1	1.9	2.1
	A	3	2	2	3	3	2	1	2.3	
S14	B	3	1	3	2	4	2	2	2.4	2.4
	A	1	2	1	1	2	1	2	1.4	
S15	B	1	1	3	1	1	1	1	1.3	1.4
	A	3	3	3	4	3	2	3	3.0	
S16	B	4	5	4	5	5	4	4	4.4	3.7
	A	2	4	3	2	3	3	2	2.7	
S17	B	3	4	3	3	4	3	1	3.0	2.9
	A	2	2	3	3	3	1	2	2.3	
S18	B	3	1	2	3	3	2	1	2.1	2.2
	A	3	3	2	3	3	3	2	2.7	
S19	B	3	4	2	3	4	4	1	3.0	2.9
	A	4	2	3	3	2	2	1	2.4	
S20	B	3	1	3	3	3	3	1	2.4	2.4
	A	3	2	1	3	3	3	1	2.3	
S21	B	4	1	1	3	4	3	1	2.4	2.4
	A	3	2	2	3	3	2	2	2.4	

Table 6 (continued)

Study ID	Researcher	QA 1	QA 2	QA 3	QA 4	QA 5	QA 6	QA 7	Score by researcher	Final score
S22	B	3	1	2	2	3	2	1	2.0	2.2
	A	4	2	2	3	3	4	2	2.9	
S23	B	3	1	3	3	3	3	1	2.4	2.6
	A	3	2	2	1	3	2	1	2.0	
S24	B	2	1	3	1	2	1	1	1.6	1.8
	A	2	2	3	2	2	3	3	2.4	
S25	B	3	1	3	1	3	3	2	2.3	2.4
	A	2	2	3	3	3	2	2	2.4	
S26	B	3	1	3	4	3	2	1	2.4	2.4
	A	3	2	2	3	3	2	3	2.6	
S27	B	3	1	3	3	3	2	4	2.7	2.6
	A	3	2	3	3	4	4	3	3.1	
S28	B	3	1	3	4	5	4	3	3.3	3.2
	A	3	2	1	3	4	4	3	2.9	
S29	B	3	1	2	4	4	5	3	3.1	3.0
	A	1	2	3	3	2	2	1	2.0	
S30	B	1	1	3	3	2	1	1	1.7	1.9
	A	2	1	3	3	2	3	1	2.1	
S31	B	1	1	3	3	1	2	1	1.7	1.9
	A	3	2	3	4	3	5	2	3.1	
S32	B	4	1	3	5	3	5	2	3.3	3.2
	A	3	2	3	4	3	3	3	3.0	
S33	B	3	1	3	4	4	4	3	3.1	3.1
	A	3	2	3	3	4	3	4	3.1	
S34	B	3	3	3	4	4	3	5	3.6	3.4
	A	3	2	2	1	3	2	2	2.1	
S35	B	3	1	2	4	4	4	1	2.7	2.4
	A	3	3	2	3	4	3	2	2.9	
S36	B	3	2	3	4	4	4	3	3.3	3.1
	A	3	2	3	4	3	4	3	3.1	
S37	B	3	1	3	3	4	4	4	3.1	3.1
	A	2	2	3	4	3	2	3	2.7	
S38	B	3	2	3	3	3	3	2	2.7	2.7
	A	3	2	3	3	3	2	3	2.7	
S39	B	3	1	4	4	3	2	3	2.9	2.8
	A	3	3	2	4	3	3	2	2.9	
S40	B	2	1	4	4	3	3	1	2.6	2.7
	A	3	2	2	3	1	1	1	1.9	
S41	B	2	1	3	3	4	3	1	2.4	2.1
	A	4	3	2	1	2	3	2	2.4	
S42	B	3	1	2	4	4	2	1	2.4	2.4

Table 7 Average quality scores of researchers by quality assessment questions

	Researcher A	Researcher B
QA1	2.8	2.7
QA2	2.1	1.4
QA3	2.4	2.7
QA4	2.8	3.0
QA5	2.9	3.4
QA6	2.6	2.8
QA7	2.1	1.9
Average	2.5	2.5

Table 8 Agreement between the researchers

Metric	Measure
Euclidean distance	14.93 (22.31%) [a]
Mean disagreement (%)	7.05
Cohen Kappa statistic (weighted)	0.42

[a] range between 0 (perfect agreement) and 66.93

5.2 Implications for research and experiments

One of the main insights extracted from this review is the difference between the number of candidate studies and selected studies. Direct exclusion criteria (such as the language or gray literature) filtered a considerable number of studies, but here we highlight the difference between the number of candidate studies after exclusion criteria and the number of selected studies.

In other words, the lack of adherence between the previously returned studies and the inclusion criteria denotes a specific property of hierarchical data stream classification: it is a roughly unexplored gap that lies between two well-known areas.

Many of the studies retrieved in this review were false positives. In some cases, the mandatory search term “hierarchical classification” was found only in the references of the retrieved study. In other cases, search terms related to data streams appear only in the background and, sometimes, related to data description (data from online systems, for example) and not to the actual classification task.

Besides, in some cases, some terms have been misused (see Sect. 1). For instance, “hierarchical classification” was used to describe a method in (Puerto-Souza et al. 2015) where classifiers were arranged in a specific order. Furthermore, other works, such as (Purohit et al. 2014) and (Huang et al. 2019), named batch classification as data stream scenarios, despite no incremental or adaptive learning is performed.

As described in Sect. 4.2, most of the proposed studies deal with images, human activity, texts, and audio. One hidden metadata in these results is that the best-performing solutions proposed by these studies tend to be specific to the domain where they were proposed, such as in (Korda et al. 2015) and (Saggese et al. 2019), where methods were developed to deal with singular problems and cannot be generalized to be applied in other areas.

This is probably due to the complexity of the proposals presented, in which it is reasonable to solve smaller and more specific problems, and then applying this knowledge to

Table 9 Adherence of the selected studies to the Hierarchical Classification of Data Stream properties

Study ID	Data Structure Υ/Ω	Label cardinality Ψ/Δ	Label Depth Φ/Ξ	Hierarchy handling Θ	Single-pass?	Readi-ness?	Bounded Memory?	Concept drift?	Time Window	Data stream handling
S1	–	–	–	–	Yes	Yes	No	No	–	Incremental
S2	–	–	–	–	–	–	–	–	–	–
S3	Tree	MPL/ MPP	PD/ NMLNP	LCN	No	No	No	No	–	Incremental
S4	Tree	SPL/ SPP	FD/ MLNP	LCN	No	No	No	No	–	Incremental
S5	Tree	SPL/ SPP	FD/ MLNP	LCPN/ LCN	No	No	No	No	Sliding	–
S6	–	–	–	–	No	Yes	No	Yes	–	Incremental
S7	–	–	–	–	Yes	Yes	Yes	Yes	Sliding	Incremental
S8	DAG - Tree	SPL/ SPP	FD/ MLNP	–	No	No	No	No	–	–
S9	Tree	MPL/ MPP	FD/ MLNP	LCPN	No	No	No	No	Sliding	–
S10	Tree	SPL/ SPP	FD/ MLNP	–	No	No	No	No	Sliding	–
S11	–	–	–	–	No	No	No	No	–	Incremental
S12	Tree	SPL/ SPP	FD/ MLNP	LCN	Yes	No	No	No	–	Incremental
S13	Tree	SPL/ SPP	FD/ MLNP	LCN	No	No	No	No	–	Incremental
S14	Tree	SPL/ SPP	FD/ MLNP	LCN	No	No	No	No	–	–
S15	Tree	SPL/ SPP	FD/ MLNP	LCN	No	No	No	No	–	–
S16	Tree	–	–	–	No	No	No	Yes	–	–
S17	Tree	MPL/ MPP	PD/ NMLNP	–	No	No	No	Yes	Landmark	Incremental
S18	Tree	SPL/ SPP	FD/ MLNP	–	No	No	No	No	–	–
S19	–	–	–	–	No	No	No	No	Sliding	–
S20	Tree	SPL/ SPP	FD/ MLNP	LCPN	No	No	No	No	Sliding	–
S21	Tree	SPL/ SPP	PD/ NMLNP	LCP	No	No	No	No	Sliding	Incremental
S22	–	–	–	–	No	No	No	No	–	Incremental
S23	–	–	–	–	No	Yes	No	No	–	–
S24	DAG - Tree	MPL/ MPP	–	–	No	No	No	Yes	–	Adaptive
S25	DAG	SPL/ SPP	FD/ MLNP	LCPN	No	No	No	No	–	–
S26	–	–	–	–	No	No	No	No	Sliding	–

Table 9 (continued)

Study ID	Data Structure Υ/Ω	Label cardinality Ψ/Δ	Label Depth Φ/Ξ	Hierarchy handling Θ	Single-pass?	Readi-ness?	Bounded Memory?	Concept drift?	Time Window	Data stream handling
S27	DAG - Tree	SPL/ SPP	FD/ MLNP	-	No	No	No	No	–	–
S28	–	–	–	–	Yes	No	No	No	–	Adaptive
S29	Tree	SPL/ SPP	FD/ MLNP	LCPN	No	No	No	No	Sliding	–
S30	Tree	SPL/ SPP	FD/ MLNP	LCPN	No	No	No	No	Sliding	–
S31	Tree	MPL/ MPP	FD/ MLNP	–	No	No	No	No	Sliding	–
S32	Tree	MPL/ MPP	FD/ MLNP	LCN	No	No	No	No	–	Incremental
S33	DAG - Tree	SPL/ SPP	FD/ MLNP	–	No	No	No	No	–	–
S34	DAG - Tree	SPL/ SPP	FD/ MLNP	LCPN	No	No	No	No	Sliding	–
S35	Tree	SPL/ SPP	FD/ MLNP	–	No	No	No	No	Sliding	–
S36	Tree	SPL/ SPP	FD/ MLNP	–	No	No	No	No	–	–
S37	–	–	–	–	Yes	Yes	Yes	Yes	Sliding	Adaptive
S38	Tree	SPL/ SPP	FD/ MLNP	–	No	No	No	No	Sliding	–
S39	Tree	SPL/ SPP	FD/ MLNP	–	No	No	No	No	–	–
S40	Tree	SPL/ SPP	FD/ MLNP	–	No	No	No	No	–	–
S41	Tree	SPL/ SPP	FD/ MLNP	–	No	No	No	No	–	–
S42	Tree	SPL/ SPP	FD/ NMLNP	LCPN	Yes	Yes	Yes	Yes	Sliding	Adaptive

more generic cases. This is also denoted by the publication venues of the selected studies, which mostly focus on the problem area rather than the computational technique itself (see Sect. 4.1), and also on commonly used evaluation algorithms and metrics, which are still based on traditional machine learning techniques.

The next step in this scenario could be the design of techniques that, from the very beginning, aggregate characteristics of both hierarchical classification and data stream classification areas together.

Another important highlight regards the datasets used in the surveyed studies. Experimental protocols have traditionally used well-known datasets in the literature as a baseline for comparing their methods. Due to the recentness of the hierarchical classification of data streams, such well-known datasets are still emerging. Not even one of the selected studies, for instance, used a dataset that had already been used by another study in the same context.

Initiatives such as UCI (Dua and Graff 2017), LibSVM (Chang and Lin 2011) and Keel (Alcalá-Fdez et al. 2011) are welcome to make datasets available in both the intersection of the areas as well as in each separately.

As possible research gaps to be investigated, this review summarized three main problems enlisted by the authors of the selected studies: non-stationarity, data complexity, and limited computational resources.

Non-stationarity concerns characteristics and classes, the consequent need for adaptation of learning models, and the detection and adaptation to concept drifts (Kotsakis et al. 2012; D'hondt et al. 2014; Peng et al. 2015; Weigl et al. 2016).

Although it seems to be a characteristic derived exclusively from data stream classification, non-stationarity also appears in hierarchical classification when we add the time component to hierarchically structured data. Data distribution can vary over time, causing classes to disappear, new classes to emerge, and relationships between classes to change. In the intra-class context, the space of discriminant characteristics of a set of examples may change over time, making models need to respond to these changes (Gomes et al. 2019).

In this sense, studies related to the dynamism of models may propose new ways to dynamically extract and select features Barddal et al. (2016, 2017), proposing the addition of emerging classes, the removal of deprecated ones in class taxonomies, and consequent problems aroused from these changes (such as class imbalance or model retraining), as well as in-depth analysis related to concept drifts, such as research concerning different types of data and time windows or statistical techniques to improve the detection of changes.

Data complexity occurs mainly in two aspects. First, considering a hierarchical structure, relationships between data need to be considered by learning models, including class taxonomies, similar concepts represented by the hierarchy, and the subsumption relation between data samples. Second, considering a data stream, data is always incomplete (as it is time-dependent) and tends to be inaccurate or at least changeable due to the way it is collected, sometimes using not monitored sensors or different types of collection.

In addition to these obstacles, there are still open problems in the area. For instance, studies related to data stream classification use the prequential method (Gama et al. 2013), considering that data labels will be available at some point in the process, which may not always be true.

In the case of adding new classes to learning models, there is still the difficulty in labeling instances linked to this new class, requiring a constant intervention of human experts to adjust the models.

Furthermore, the authors in (Djorgovski et al. 2012) and (D'hondt et al. 2014) report difficulties in considering contextual information in their models, raising the question of how learning models may be able to integrate data and information related to their prediction task into their processes and which are not present in the models, but in correlated datasets with different and unpredictable formats.

Finally, the authors in (Wang et al. 2009; Song et al. 2009, 2011; Pesaranghader et al. 2018) report as a research gap studies that consider the computational resources involved in classification tasks, which was highlighted by the large-scale data used as input in the data stream classification process and reinforced by the additional complexity brought from the class hierarchy, which already usually works with a high number of classes organized in different hierarchical levels.

There is room for studies regarding the hierarchical classification of data streams that comprehend methods able to work with bounded memory and to adapt their models without the need for retraining. It is important to note that there is even a need for computational cost analysis in previously published methods.

We suggest that new studies always present computational performance analysis considering at least memory and time consumption and system readiness (to be ready to predict at any time), which is already common in data stream classification studies.

It is also important to authors of new studies to be careful about proposing methods tested only under some conditions that may not perform well in the actual environment for which it was designed; sometimes, such environment is limited in computational resources or even in the possibility of data collection, making the proposed method impracticable.

6 Concluding Remarks

This systematic review comprised hierarchical classification of data streams, i.e., an intersection area between two well-known classification tasks: hierarchical classification, with data organized in classes with parent and child nodes, and data stream classification, with large-scale and non-stationary data.

We investigated the topic under five main research questions related to the problems (`RQ1`) handled by hierarchical classification of data streams, the datasets (`RQ2`) used in the studies and experiments, the algorithms (`RQ3`) and evaluation metrics (`RQ4`) used, and the research gaps (`RQ5`) and possible subjects of further research.

We have conducted a literature search and identified 42 primary original studies comprising the five research questions that guided this review. The main findings of this review are summarized as follows.

- **RQ1**: The main problems handled by hierarchical classification of data streams include classification of images, human activities, texts, and audio.
- **RQ2**: The datasets used in the study experiments are mostly created for the study or comprise synthetic data, revealing that there is not yet a baseline for the creation of testbeds.
- **RQ3**: The used algorithms are mainly well-known techniques, such as Support Vector Machines, K-Nearest Neighbors or Neural Networks, or adaptations of those.

- **RQ4**: The used evaluation metrics are mainly well-known techniques, such as accuracy, precision, recall, or f-measure, with the addition of the Loss metric and other metrics concerned with computational resources.
- **RQ5**: Research gaps in the hierarchical classification of data streams context reported by the authors are mainly related to dynamism (ever-changing environment), data complexity (including large-scale data and non-stationary data over time), and computational resources (such as bounded memory and hardware limitations in real-world applications).

This review also provided implications for research and experiments. We recommend that new studies comprise methods that consider characteristics of the hierarchical classification and data stream classification together since the conceptual design of the method. Also, we believe that it is important for new studies to evidence their experiment protocols, and mainly their datasets, to allow other studies to perform further analyses and create comparison baselines in the area.

Furthermore, we may consider that problems previously mapped in the single areas (such as dynamism in classification or computational resources used) need to be revisited from the perspective of the hierarchical classification of data streams and new studies need to always take computational resource usage into account since data streams are potentially unbounded.

As limitations, we point out that our analysis of the problems dealt with by the studies comprised the main category of the problem, disregarding other characteristics or subcategories. Regarding datasets, the characteristics of the data or their distribution (such as class imbalance and the existence of concept drifts) were outside the scope of this review. Also, this review did not intend to describe how the algorithms and evaluation metrics used by the selected studies work or to compare them.

When assessing the quality of the selected studies, we tried to minimize threats to the reliability and biases in the evaluation using the same criteria for all the studies. Nevertheless, one must be mindful while using these results to draw inferences. Also, the quality scores should not be used to rank the studies.

We also highlight that, concerning our results, it is not possible to specify which of the findings of this review are exclusive of the hierarchical classification of data streams since there are minorno defined characteristics already existent in hierarchical classification and data stream classification areas separately.

Thus, for future work, new systematic literature reviews can be considered to investigate the two base areas (hierarchical classification and data stream classification) of this review separately and then clarify which features found in this review are also found in the base areas (so are related to the classification task in general) and which ones are exclusive of the hierarchical classification of data stream task.

Appendix: Datasets, algorithms and evaluation metrics used in the selected studies and research questions addressed

See Table 10.

Table 10 Datasets, algorithms and evaluation metrics used in the selected studies and research questions addressed

Study ID	Dataset type	Dataset name/description	Algorithm(s) used	Evaluation metric(s) used	RQs addressed
S1	Obtained from literature	LED e Waveform (UCI).	Incremental Algorithm Driven by Error Margins (IADEM), Decision Tree and Naive Bayes.	Induction time, tree size and accuracy.	1, 2, 3, 4, 5
S2	–	–	–	–	–
S3	Created to the study	Webpages obtained using crawlers.	m-sv-incremental.	F-measure and h-Loss.	1, 2, 3, 4, 5
S4	Synthetic	Simulated user activities obtained using IBM Websphere Workload Simulator.	Perceptron.	Accuracy and training time consumption.	1, 2, 3, 4
S5	Synthetic	Simulated signal obtained using Monte Carlo Simulation.	Multi-layer Perceptron and Kernel density estimators.	Accuracy and rejection capability.	1, 2, 3, 4, 5
S6	Created to the study	Data from a Chemical Vapor Deposition (CVD) equipment.	Principal component analysis, Genetic algorithm and Support Vector Machine.	Accuracy and classification error comparison.	1, 2, 3, 4, 5
S7	Created to the study	Data from images of CD imprint inspection, egg inspection and inspections of metal rotor parts.	Evolving vector quantization (eVQ-Class) and FLEXible Fuzzy Inference Systems for classification (FLEXFIS-Class).	Accuracy (total and over number of instances).	1, 2, 3, 4, 5
S8	Obtained from literature	Worldwide Incidents Tracking System (WITS) e MIPT Terrorism Knowledge Base.	Fuzzy Formal Concept Analysis.	–	1, 2, 3
S9	Created to the study	CCTV Broadcast News.	SVM-BinaryTree.	Accuracy.	1, 2, 3, 4
S10	Created to the study	Network traffic of video conference.	Growing Hierarchical Self-Organizing Map (GHSOM).	Accuracy.	1, 2, 3, 4
S11	Obtained from literature	Caltech 101 database.	Hierarchical generative model (Hierarchical GM).	Accuracy (per class).	1, 2, 3, 4, 5
S12	Obtained from literature	Catalina Real-time Transient Survey (CRTS).	Naive Bayes, Decision Tree and Markov Logic Networks.	Accuracy.	1, 2, 3, 4, 5
S13	Synthetic	Simulated user activities obtained using IBM Websphere Workload Simulator.	Perceptron.	F-measure.	1, 2, 3, 4, 5

Table 10 (continued)

Study ID	Dataset type	Dataset name/description	Algorithm(s) used	Evaluation metric(s) used	RQs addressed
S14	Created to the study	Greek radio programme obtained via podcast.	Multi-layer Perceptron, Nearest Neighbors, Logistic Regression, Decision Tree (J48 and Tree Complex) and Sequential Minimal Optimization.	Accuracy.	1, 2, 3, 4, 5
S15	Synthetic	Data from temperature, humidity and light intensity sensors.	XMLBoost.	Accuracy, time consumption and convergence rate.	1, 2, 3, 4, 5
S16	Obtained from literature	CLEF-IP 2011 patent corpus.	Balanced Winnow.	F-measure (micro and macro-averaged)	1, 2, 3, 4, 5
S17	Created to the study	Twitter Data (search-based API).	structural Support Vector Machines (C-SSVM, NDC-SSVM, GTC-SSVM and LTC-SSVM), CSSA (Kernel density estimation based HMC method), CLUS-HMC (Decision tree-based HMC method), H-SVM (Hierarchical SVM for multi-label classification), CSHC (Structural multi-class learning method) and NBC (Naive Bayesian method).	Accuracy and F-measure.	1, 2, 3, 4
S18	Created to the study	Twitter Data (search-based API).	Natural Language Processing and Decision Tree.	Precision, Recall and Kappa.	1, 2, 3, 4
S19	Created to the study	Data from human–human interaction collected using a Kinect sensor.	Support Vector Machine.	Precision, Recall and F-measure.	1, 2, 3, 4
S20	Obtained from literature	ASPIS (Athens Study for Psychosis Proneness and Incidence of Schizophrenia).	Artificial Neural Networks.	Sensitivity, Specificity, Accuracy and Receiver Operating Curve.	1, 2, 3, 4
S21	Obtained from literature	ILSVRC2010 imageset and NUS-WIDE2 imageset.	structural Support Vector Machine.	Accuracy and computational efficiency.	1, 2, 3, 4
S22	Created to the study	Restaurant objects obtained by acquiring scene images.	Nearest-Neighbors	F-measure.	1, 2, 3, 4

Table 10 (continued)

Study ID	Dataset type	Dataset name/description	Algorithm(s) used	Evaluation metric(s) used	RQs addressed
S23	Created to the study	Twitter (streaming-based API).	Support Vector Machine.	F-measure and detection time.	1, 2, 3, 4, 5
S24	–	–	Semantic Hierarchical multi-label classification.	F-measure.	3, 4
S25	Created to the study	Underwater data of marine life from Autonomous Underwater Vehicles (AUV).	Haar feature-based cascade classifier and Multi-Descriptor Random Forests (MD-NCMFs).	F-measure.	1, 2, 3, 4
S26	Created to the study	Frames extracted from colonoscopy videos from Hospital Universitario del Valle Evaristo Garcia (Colombia).	Support Vector Machine and Hierarchical Classifier with Temporal Consistency.	Sensitivity, Specificity, and Mean Class Accuracy.	1, 2, 3, 4
S27	Created to the study	Human activity data obtained using smartphone built-in sensors.	Holistic triple learning (HTL), Support Vector Machine and probability.	Accuracy e energy consumption.	1, 2, 3, 4
S28	Created to the study	Images of microfluidic chip and sheaves obtained from inspection systems.	Random Forests e Evolving fuzzy classifiers.	Accuracy.	1, 2, 3, 4, 5
S29	Obtained from literature	Audio files collected from Google translate, Oxford and Merriam-Webster dictionaries.	K-Nearest Neighbors (with Dynamic time warping distance).	Accuracy.	1, 2, 3, 4, 5
S30	Created to the study	Human activity data obtained using a wristband mounted system.	Random Forests, Support Vector Machine, Naive Bayes, K-Nearest Neighbors, Boosting, Bagging, Plurality Voting and Stacking.	Accuracy.	1, 2, 3, 4
S31	Obtained from literature	Pascal Dataset.	Sliding Window Stochastic Gradient Descent Algorithm (SW-SGD)[a].	Loss.	2, 3, 4

Table 10 (continued)

Study ID	Dataset type	Dataset name/description	Algorithm(s) used	Evaluation metric(s) used	RQs addressed
S32	Obtained from literature	Dermatology, Flare, Forest, Frogs, Glass, Optdigits, Pendigits, SatImage, Segmentation, Sports, Texture, TrafficLight, Vehicle, Vertebral, Vowel and Zoo. (UCI, LibSVM and Kell)	Hierarchical Multiclass Classification with Rafraiming Approach (HMC-RA), Decision Tree (C50), K-Nearest Neighbors, Recursive partitioning tree , Neural Network (NNET), Random Forest and Support Vector Machine.	Accuracy and time consumption.	2, 3, 4, 5
S33	Obtained from literature	FABO dataset.	Deep Network (Convolutional neural networks, Bilateral long short-term memory recurrent neural networks (BLSTM-RNN) and Principal component analysis).	Macro Average Accuracy (MAA) and Accuracy.	1, 2, 3, 4
S34	Obtained from literature	PAMAP2 dataset.	Decision Tree, Support Vector Machine, Random Forests, AdaBoost, and Extreme learning machines (ELM).	Accuracy.	1, 2, 3, 4
S35	Obtained from literature	FIGER and OntoNotes.	Natural Language Processing and Neural Networks.	Micro F-measure.	1, 2, 3, 4
S36	Obtained from literature	HAR (Human Activity Recognition) (UCI)	Group-based Context-aware method for human activity recognition on smartphones (GCHAR), Random Tree (RT), Decision Table, Decision Tree (J48), BayesNet, K-Nearest Neighbors and Bagging.	Accuracy, F-measure and training time.	1, 2, 3, 4
S37	Obtained from literature	Sine1, Sine2, Mixed, Stagger, Circles, LED (Synthetic); Electricity, Forest CoverType, Poker hand (Real-world) ; Adult, Nursery, Shuttle (UCI).	Naive Bayes, Decision Stump, Hoeffding Tree, Perceptron, K-Nearest Neighbors and Stacking Fast Hoeffding Drift Detection Method (FHDDMS and FHDDMSadd)[a].	CAR (Classification, Adaptation, and Resource Consumption).	2, 3, 4, 5
S38	Created to the study	MPEG-7 data from a video.	Deep Convolutional Neural Network.	Pearson product-moment correlation.	1, 2, 3, 4

Table 10 (continued)

Study ID	Dataset type	Dataset name/description	Algorithm(s) used	Evaluation metric(s) used	RQs addressed
S39	Created to the study	Air quality data obtained from crowdsourcing vehicles with air quality sensing devices.	Random Forests.	Recall.	1, 2, 3, 4
S40	Obtained from literature	OASIS and Alzheimer's disease neuroimaging initiative (ADNI).	Deep Convolutional Neural Network and Support Vector Machine.	Accuracy, Precision, F-measure, Sensitivity and Specificity.	1, 2, 3, 4
S41	Obtained from literature	MIVIA-Skeleton (MIVIA-S), MSR Daily Activity 3D (MSRDA) and Berkeley Multimodal Human Action Database (MHAD).	Fast global alignment kernel (string kernels), Support Vector Machine and K-Nearest Neighbors.	Recognition Rate (RR) and Error Rate (ER)e miss rate (MR).	1, 2, 3, 4
S42	Created to the study	Entomology, Ichthyology, and Audio processing datasets.	Incremental K-Nearest Neighbors.	Hierarchical Precision (hP), Hierarchical Recall (hR) and Hierarchical F-measure (hF).	1, 2, 3, 4, 5

[a] These algorithms are not learning algorithms but related to the classification task

Author Contributions Not applicable.

Funding This study was partially financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Data Availability Not applicable.

Code Availability Not applicable.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Aha DW, Kibler D, Albert MK (1991) Instance-based learning algorithms. Machine learn 6(1):37–66

Alazrai R, Mowafi Y, Lee CG (2015) Anatomical-plane-based representation for human-human interactions analysis. Pattern Recogn 48(8):2346–2363

Alcalá-Fdez J, Fernández A, Luengo J, Derrac J, García S, Sánchez L, Herrera F (2011) Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. J Multiple-Valued Logic Soft Comput 17:1

Anderez DO, Appiah K, Lotfi A, Langesiepen C (2017) A hierarchical approach towards activity recognition. In: Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments, ACM, pp 269–274

Atkins S, Lewin S, Smith H, Engel M, Fretheim A, Volmink J (2008) Conducting a meta-ethnography of qualitative literature: lessons learnt. BMC Med Res Methodol 8(1):21

Babcock B, Babu S, Datar M, Motwani R, Widom J (2002) Models and issues in data stream systems. In: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, ACM, pp 1–16

Barddal JP, Gomes HM, Enembreck F, Pfahringer B, Bifet A (2016) On dynamic feature weighting for feature drifting data streams. In: Joint european conference on machine learning and knowledge discovery in databases, Springer, pp 129–144

Barddal JP, Gomes HM, Enembreck F, Pfahringer B (2017) A survey on feature drift adaptation: definition, benchmark, challenges and future directions. J Syst Softw 127:278–294

Bifet A, Kirkby R (2009) Data stream mining a practical approach

Breiman L (2001) Random forests. Machine learn 45(1):5–32

Burred JJ, Lerch A (2003) A hierarchical approach to automatic musical genre classification. In: Proceedings of the 6th international conference on digital audio effects, Citeseer, pp 8–11

Cano A (2018) A survey on graphic processing unit computing for large-scale data mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8(1):e1232

Cao L, Wang Y, Zhang B, Jin Q, Vasilakos AV (2018) Gchar: an efficient group-based context–aware human activity recognition on smartphone. J Parallel Distrib Comput 118:67–80

Cerri R, Pappa GL, Carvalho ACP, Freitas AA (2015) An extensive evaluation of decision tree-based hierarchical multilabel classification methods and performance measures. Comput Intell 31(1):1–46

Chakroun I, Haber T, Ashby TJ (2017) Sw-sgd: The sliding window stochastic gradient descent algorithm. Procedia Computer Science 108:2318–2322 https://doi.org/10.1016/j.procs.2017.05.082, http://www.sciencedirect.com/science/article/pii/S1877050917306221, International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland

Chang CC, Lin CJ (2011) LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2:27:1–27:27, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Chavez AG, Fontes J, Afonso P, Pfingsthorn M, Birk A (2016) Automated species counting using a hierarchical classification approach with haar cascades and multi-descriptor random forests. In: OCEANS 2016-Shanghai, IEEE, pp 1–6

Chen Z, Wu J, Castiglione A, Wu W (2016) Human continuous activity recognition based on energy-efficient schemes considering cloud security technology. Security Commun Net 9(16):3585–3601

Chou PH, Wu MJ, Chen KK (2010) Integrating support vector machine and genetic algorithm to implement dynamic wafer quality prediction system. Expert Syst Appl 37(6):4413–4424

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Measur 20(1):37–46

del Campo-Ávila J, Ramos-Jiménez G, Gama J, Morales-Bueno R (2008) Improving the performance of an incremental algorithm driven by error margins. Intell Data Analy 12(3):305–318

Defiyanti S, Winarko E, Priyanta S (2019) A survey of hierarchical classification algorithms with big-bang approach. In: 2019 5th International Conference on Science and Technology (ICST), IEEE, vol 1, pp 1–6

Djorgovski SG, Mahabal A, Donalek C, Graham MJ, Drake AJ, Moghaddam B, Turmon M (2012) Flashes in a star stream: Automated classification of astronomical transient events. arXiv preprint arXiv:12091681

Domingos P, Hulten G (2000) Mining high-speed data streams. In: Kdd, vol 2, p 4

Dua D, Graff C (2017) UCI machine learning repository. http://archive.ics.uci.edu/ml

Dumais S, Chen H (2000) Hierarchical classification of web content. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp 256–263

D'hondt E, Verberne S, Oostdijk N, Beney J, Koster C, Boves L, (2014) Dealing with temporal variation in patent categorization. Inf Retrieval 17(5–6):520–544

Fan J, Zhang J, Mei K, Peng J, Gao L (2015) Cost-sensitive learning of hierarchical tree classifiers for large-scale image classification and novel category detection. Pattern Recogn 48(5):1673–1687

Freitas A, Carvalho A (2007) A tutorial on hierarchical classification with applications in bioinformatics. In: Research and trends in data mining technologies and applications, IGI Global, pp 175–208

Gaber MM, Zaslavsky A, Krishnaswamy S (2005) Mining data streams: a review. ACM SIGMOD Rec 34(2):18–26

Gama J (2010) Knowledge discovery from data streams. Chapman and Hall/CRC

Gama J, Sebastião R, Rodrigues PP (2013) On evaluating stream learning algorithms. Mach Learn 90(3):317–346

Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A (2014) A survey on concept drift adaptation. ACM Comput Surv (CSUR) 46(4):44

Gomes HM, Read J, Bifet A, Barddal JP, Gama J (2019) Machine learning for streaming data: state of the art, challenges, and opportunities. ACM SIGKDD Explorations Newsl 21(2):6–22

Gu P, Qs Zhu, Zhang C, Zhuang Z (2009) An adaptive hierarchical model based on fusion of ontology and context. Transac Beijing Instit Technol 10:1

Hamooni H, Mueen A, Neel A (2016) Phoneme sequence recognition via dtw-based classification. Knowl Inf Syst 48(2):253–275

Higgins JP, Green S (2011) Cochrane handbook for systematic reviews of interventions, vol 4. Wiley

Huang J, Duan N, Ji P, Ma C, Ding Y, Yu Y, Zhou Q, Sun W et al (2018) A crowdsource-based sensing system for monitoring fine-grained air quality in urban environments. IEEE Internet Things J 6(2):3240–3247

Huang KY, Wu CH, Hong QB, Su MH, Chen YH (2019) Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 5866–5870

Jeong S, Lee M (2012) Adaptive object recognition model using incremental feature representation and hierarchical classification. Neural Netw 25:130–140

Kasaei SH, Oliveira M, Lim GH, Lopes LS, Tomé AM (2015) Interactive open-ended learning for 3d object recognition: an approach and experiments. J Intell Robotic Syst 80(3–4):537–553

Kauppi JP, Martikainen K, Ruotsalainen U (2010) Hierarchical classification of dynamically varying radar pulse repetition interval modulation patterns. Neural Netw 23(10):1226–1237

Khowaja SA, Prabono AG, Setiawan F, Yahya BN, Lee SL (2018) Contextual activity based healthcare internet of things, services, and people (hiotsp): an architectural framework for healthcare monitoring using wearable sensors. Comput Netw 145:190–206

Kiritchenko S, Famili F (2005) Functional annotation of genes using hierarchical text categorization. Proceedings of BioLink SIG, ISMB

Kitchenham B, Charters S (2007) Guidelines for performing systematic literature reviews in software engineering

Korda AI, Asvestas PA, Matsopoulos GK, Ventouras EM, Smyrnis NP (2015) Automatic identification of oculomotor behavior using pattern recognition techniques. Comput Biol Med 60:151–162

Kosmopoulos A, Partalas I, Gaussier E, Paliouras G, Androutsopoulos I (2015) Evaluation measures for hierarchical classification: a unified view and novel approaches. Data Min Knowl Disc 29(3):820–865

Kotsakis R, Kalliris G, Dimoulas C (2012) Investigation of broadcast-audio semantic analysis scenarios employing radio-programme-adaptive pattern classification. Speech Commun 54(6):743–762

Krempl G, Žliobaite I, Brzeziński D, Hüllermeier E, Last M, Lemaire V, Noack T, Shaker A, Sievi S, Spiliopoulou M et al (2014) Open challenges for data stream mining research. ACM SIGKDD Explorations Newsl 16(1):1–10

La L, Guo Q, Alonso L, Zhang F (2014) Classifying xml data of semantic sensor networks. Arab J Sci Eng 39(5):3733–3745

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. biometrics ,1: 159–174

Liu J, Wang L, Zhou M, Wang J, Lee S (2018) Fine-grained entity type classification with adaptive context. Soft Comput 22(13):4307–4318

Lu Y (1997) Concept hierarchy in data mining: Specification, generation and implementation. PhD thesis, Theses (School of Computing Science)/Simon Fraser University

Lughofer E (2010) On-line evolving image classifiers and their application to surface inspection. Image Vis Comput 28(7):1065–1079

Martin T, Shen Y, Majidian A (2010) Soft concept hierarchies to summarise data streams and highlight anomalous changes. In: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Springer, pp 44–54

Masud MM, Chen Q, Khan L, Aggarwal C, Gao J, Han J, Thuraisingham B (2010) Addressing concept-evolution in concept-drifting data streams. In: 2010 IEEE International Conference on Data Mining, IEEE, pp 929–934

Melo A, Völker J, Paulheim H (2017) Type prediction in noisy rdf knowledge bases using hierarchical multilabel classification with graph and latent features. Int J Artif Intell Tools 26(02):1760011

Mermillod M, Bugaiska A, Bonin P (2013) The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects. Front Psychol 4:504. https://doi.org/10.3389/fpsyg.2013.00504, http://journal.frontiersin.org/article/10.3389/fpsyg.2013.00504/abstract

Nguyen HL, Woon YK, Ng WK, Wan L (2012) Heterogeneous ensemble for feature drifts in data streams. In: Pacific-Asia conference on knowledge discovery and data mining, Springer, pp 1–12

Nguyen HL, Woon YK, Ng WK (2015) A survey on data stream clustering and classification. Knowl Inf Syst 45(3):535–569

Noblit GW, Hare RD (1988) Meta-ethnography: Synthesizing qualitative studies, vol 11. sage

Parmezan ARS, Souza VM, Batista GE (2018) Towards hierarchical classification of data streams. In: Iberoamerican Congress on Pattern Recognition, Springer, pp 314–322

Peixoto R, Cruz C, Silva N (2016) Adaptive learning process for the evolution of ontology-described classification model in big data context. In: 2016 SAI Computing Conference (SAI), IEEE, pp 532–540

Peng B, Li J, Chen J, Han X, Xu R, Wong KF (2015) Trending sentiment-topic detection on twitter. In: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, pp 66–77

Pereira RM, Costa YM, Silla CN (2021) Handling imbalance in hierarchical classification problems using local classifiers approaches. Data Mining and Knowledge Discovery pp 1–58

Pesaranghader A, Viktor H, Paquet E (2018) Reservoir of diverse adaptive learners and stacking fast hoeffding drift detection methods for evolving data streams. Mach Learn 107(11):1711–1743

Protasov S, Khan AM, Sozykin K, Ahmad M (2018) Using deep features for video scene detection and annotation. SIViP 12(5):991–999

Puerto-Souza GA, Manivannan S, Trujillo MP, Hoyos JA, Trucco E, Mariottini GL (2015) Enhancing normal-abnormal classification accuracy in colonoscopy videos via temporal consistency. In: Computer-Assisted and Robotic Endoscopy, Springer, pp 129–139

Purohit H, Hampton A, Bhatt S, Shalin VL, Sheth AP, Flach JM (2014) Identifying seekers and suppliers in social media communities to support crisis coordination. Comput Supported Coop Work (CSCW) 23(4–6):513–545

Quinlan JR (2014) C4. 5: Programs for Machine Learning. Elsevier

Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND (2009) Dataset shift in machine learning. The MIT Press

Ramírez-Gallego S, Krawczyk B, García S, Woźniak M, Herrera F (2017) A survey on data preprocessing for data stream mining: current status and future directions. Neurocomputing 239:39–57

Raza M, Awais M, Ellahi W, Aslam N, Nguyen HX, Le-Minh H (2019) Diagnosis and monitoring of alzheimer's patients using classical and deep learning techniques. Expert Syst Appl

Ren Z, Peetz MH, Liang S, Van Dolen W, De Rijke M (2014) Hierarchical multi-label classification of social text streams. In: Proceedings of the 37th International ACM SIGIR Conference on Research & development in Information Retrieval, ACM, pp 213–222

Saggese A, Strisciuglio N, Vento M, Petkov N (2019) Learning skeleton representations for human action recognition. Pattern Recogn Lett 118:23–31

Shi H, Hamagami T, Xu H, Yu P, Wu Y (2012) A method for classifying packets into network flows based on ghsom. Mobile Netw Appl 17(6):730–739

Silla CN, Freitas AA (2011) A survey of hierarchical classification across different application domains. Data Min Knowl Disc 22(1–2):31–72

Silva-Palacios D, Ferri C, Ramirez-Quintana MJ (2018) Adapting hierarchical multiclass classification to changes in the target concept. In: Conference of the Spanish Association for Artificial Intelligence, Springer, pp 118–127

Song Y, Sailer A, Shaikh H (2009) Problem classification method to enhance the itil incident and problem. In: 2009 IFIP/IEEE International Symposium on Integrated Network Management, IEEE, pp 295–298

Song Y, Sailer A, Shaikh H (2011) Hierarchical online problem classification for it support services. IEEE Trans Serv Comput 5(3):345–357

Sun B, Cao S, He J, Yu L (2018) Affect recognition from facial movements and body gestures by hierarchical deep spatio-temporal features and fusion strategy. Neural Netw 105:36–51

Tsymbal A (2004) The problem of concept drift: definitions and related work. Comput Sci Depart Trinity College Dublin 106(2):58

Venkatesan R, Er MJ (2014) Multi-label classification method based on extreme learning machines. In: 2014 13th International Conference on Control Automation Robotics & Vision (ICARCV), IEEE, pp 619–624

Wang Y, Gong Z, Guo J (2009) Hierarchical classification of business information on the web using incremental learning. In: 2009 IEEE International Conference on e-Business Engineering, IEEE, pp 303–309

Weigl E, Heidl W, Lughofer E, Radauer T, Eitzinger C (2016) On improving performance of surface inspection systems by online active learning and flexible classifier updates. Mach Vis Appl 27(1):103–127

Wen J, Li S, Lin Z, Hu Y, Huang C (2012) Systematic literature review of machine learning based software development effort estimation models. Inf Softw Technol 54(1):41–59

Widmer G, Kubat M (1996) Learning in the presence of concept drift and hidden contexts. Mach Learn 23(1):69–101

Wu F, Zhang J, Honavar V (2005) Learning classifiers using hierarchically structured class taxonomies. In: International Symposium on Abstraction, Reformulation, and Approximation, Springer, pp 313–320

Xie L, Fu ZH, Feng W, Luo Y (2011) Pitch-density-based features and an svm binary tree approach for multi-class audio classification in broadcast news. Multimedia Syst 17(2):101–112

Yassin NI, Omran S, El Houby EM, Allam H (2018) Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. Comput Methods Programs Biomed 156:25–45

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