Artificial Intelligence Review (2024) 57:228
https://doi.org/10.1007/s10462-024-10845-9

Three-way decision in machine learning tasks: a systematic review

Andrea Campagner[1] · Frida Milella[2] · Davide Ciucci[2] · Federico Cabitza[1,2]

Accepted: 24 June 2024 / Published online: 2 August 2024

Abstract
In this article, we survey the applications of Three-way decision theory (TWD) in machine learning (ML), focusing in particular on four tasks: weakly supervised learning and multi-source data management, missing data management, uncertainty quantification in classification, and uncertainty quantification in clustering. For each of these four tasks we present the results of a systematic review of the literature, by which we report on the main characteristics of the current state of the art, as well as on the quality of reporting and reproducibility level of the works found in the literature. To this aim, we discuss the main benefits, limitations and issues found in the reviewed articles, and we give clear indications and directions for quality improvement that are informed by validation, reporting, and reproducibility standards, guidelines and best practice that have recently emerged in the ML field. Finally, we discuss about the more promising and relevant directions for future research in regard to TWD.

Keywords Three-way decision · Machine learning · Artificial intelligence · Systematic literature review

1 Introduction

Three-way decision (TWD) is an emerging conceptual and computational paradigm to represent, handle and process uncertainty inspired by rough set theory (Pawlak 1982, 1991) , which was originally proposed by Yao (2010, 2012) . Intuitively and in its most general and abstract form, TWD is based on the idea of approaching computational problem-solving from a *ternary*, rather than binary, perspective. In this setting, binary perspective refers to computational processes that are based on the act of discriminating the objects of interests into those that satisfy a set of desirable requirements and those that do not. Instead, the ternary perspective adopted by TWD grounds on a *tripartition* of the universe of interest, where also a third category is also considered associated with objects whose status is *uncertain*. This ternary perspective is conceptually based on the

✉ Andrea Campagner
andrea.campagner@unimib.it

1 IRCCS Istituto Ortopedico Galeazzi, Milan, Italy

2 University of Milano-Bicocca, Milan, Italy

trisecting-acting-outcome (TAO) model (Yao 2018) : all computational processes that are involved in *Trisecting* have the objective of dividing the universe under investigation into three partitions in order to distinguish certain objects from uncertain ones, i.e. to employ the above described tripartitioning of the universe of interest; *Acting* describes the computational steps dealing with the three parts identified that specifically make explicit how to manage the uncertain objects that have been previously identified; and *Outcome* provides methodology for evaluating the results obtained as well as, by extension, the methodology employed.

From a theoretical point of view, the above mentioned general ideas can be formalized through a set-theoretic approach. Namely, we assume the existence of a universal set U of objects of interest: for example, in the Machine Learning setting (which will be our main focus within this article), U can represent the set of all potential instances for a given task or problem. The fundamental idea in TWD is the introduction of a trisecting function $\pi : U \rightarrow \{P, N, Bnd\}$ that distinguishes elements of U into *certain* (i.e., elements in $P \cup N$) and *uncertain* ones (i.e., elements in Bnd). Such a trisecting function can be implemented in many different ways, depending on the considered application. The acting and outcome steps described above can, on the other hand, be formalized in terms of, respectively, computational procedures (i.e., algorithms) and metrics that allow us to process the results of π, and finally evaluate the results of such data processing steps. We will provide additional details on these two steps in the following, when we will focus on some specific applications of TWD in Machine Learning.

Among many several applications in the data sciences (Ma 2016; Yang and Hou 2018; Yao 2022) , the application of TWD as a general-purpose framework to handle uncertainty in Machine Learning (ML) has attracted particular interest in the recent years. In order to systematize the contributions to the application and development of TWD in ML, in a recent narrative survey (Campagner et al. 2020a) we proposed a categorization of applications of TWD-based approaches in ML, which distinguishes between methods that deal with uncertainty in either the input or the output of a Machine Learning pipeline (Hapke and Nelson 2020), as shown in Fig. 1. In both cases, the above mentioned conceptual framework underlying TWD can be specified by defining the universal set U to be a space of instances, described in terms of some feature space X (usually taken to be a n-dimensional real vector space $\mathbb{R}^n$ or, more generally, a n-dimensional set of symbolic and numerical characteristics), which encode relevant information about the instances which is deemed to be useful for their

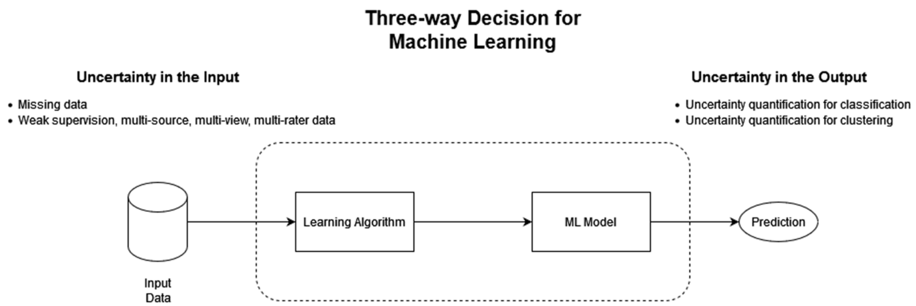

Fig. 1 A graphical representation of the framework adopted in this article to classify applications of Three-way decision in Machine Learning

categorization or analysis, as well as (at least in classification tasks) a target space Y, which encodes the categories relevant for the considered application task.

The term *Uncertainty in the input* denotes tasks and problems in which the training datasets utilized by ML learning algorithms contain explicit instances of uncertainty (Destercke 2022). In these cases, the trisecting function π typically acts as a way to separate certain instances (which can be directly manipulated using traditional ML tools) from uncertain ones (which, by contrast, require some further processing). Following Campagner et al. (2020a), we distinguish the uncertainty in the input in two common variants: *Incomplete data* and *weak or multi-source supervision*. Incomplete data refers to settings in which some predictive features' values in the dataset are missing (Little and Rubin 2019) or are otherwise incomplete (Miao et al. 2018; Williams et al. 2007) . Formally speaking, this amounts to assuming that, in the above described representation of instances in terms of features and targets, some information about the feature space X could be missing: as an example, if we are given an instance $x \in X = \mathbb{R}^n$, then some of the n features of x could be unknown. Hence, in this setting, π aims at separating complete instances from incomplete data, so that the latter ones can be properly managed through the acting step (Emmanuel et al. 2021) : common implementations of this acting step could involve discarding the incomplete data, filling in the missing information, or using ML techniques that are somehow able to directly use such faulty instances. Weak and multi-source supervision, on the other hand, refers to settings in which the incompleteness affects the supervision (i.e. the target or decision variable) or the relation between the predictive features and the supervision itself, which are then only partially specified. Thus, formally speaking, the case of weak supervision can be seen as the dual of incomplete data, where the incompleteness affects the target space Y rather than the feature space X: then, the trisecting, acting and outcome steps could be seen as a more or less direct adaptation of their instantiation in the case of incomplete data. While, similarly to the case of incomplete data, weak and multi-source supervision can arise in a variety of natural scenarios (Campagner et al. 2021b; Lienen and Hüllermeier 2021; Poyiadzi et al. 2022) , this sort of uncertainty has only recently garnered a growing amount of attention (Poyiadzi et al. 2022; Zhou 2018) , especially motivated by the data acquisition bottleneck associated with the big data requirements for modern ML models and the related growth in multi-rater, multi-source and multi-view data acquisition practices.

By contrast, *uncertainty in the output* characterizes the so-called cautious AI. Recently, this expression has been proposed (Campagner et al. 2021; Hüllermeier and Waegeman 2021) to denote AI applications that expose their uncertainty quantification[1] and, according to this latter, might reject the user request for a single-value, clear-cut classification of any new instance and instead provide a more uncertainty-informed advice about the case. In other words, uncertainty quantification mechanisms are aimed at making ML models more "robust" by making its (inherent and partly insuppressible) predictive uncertainty more explicit and allowing the model to partially abstain. Formally speaking, this means that the uncertainty in the tasks (and hence the object of the trisecting function π) regards not so much the universal set U in and by itself, but rather how this is processed and *seen through the lens of* a ML model M: such a model is constructed by applying some *learning algorithm A* to a subset of data drawn from our universal set U, and is usually intended as a way to reconstruct some desired pattern or characteristics of interest from such a finite amount

[1] With uncertainty quantification we here mean both uncertainty representation and estimation.

of data, in such a manner that this categorization could then be reproduced on new data drawn from U. The application of TWD to implement uncertainty quantification methods in ML has been one of the main aims of TWD since its origin, both for *classification* , where TWD has been originally applied to spam detection (Yao 2010; Zhou et al. 2010), and for *clustering* , where TWD-based clustering (Yu 2017) emerged as an offspring of rough clustering (Lingras and West 2004) and interval set clustering (Yao et al. 2009) . As is the case in both settings , the need for uncertainty quantification arises from the fact that a ML model may not be able to separate or discriminate instances, and thus may fail at assigning a precise, single label to them. Such a situation may arise for a variety of reasons: the considered set of features may not be large or informative enough; the selected ML model may not be sufficiently expressive to solve the task of interest; the instances at hand may lie on the boundary of the decision space and hence may be "too close" to similar, yet differently classified, instances[2]. Allowing the classifier or clusterer to abstain (Yao 2012), even partially, which means discarding some of the potential alternative classifications, is the strategy advised by TWD in this setting. In such a process, according to the TAO model described above, the focus is on the trisecting function[3] π , which involves determining which instances the machine-learning model M (regardless of whether it is used for classification or clustering) should regard as uncertain, and accordingly abstain from making predictions. In this sense, the acting step for tasks related to uncertainty in the output simply amounts to confirming the results of the trisecting function π: that is, if the instance x is deemed to be certain (i.e., $\pi(x) \notin Bnd$), then the prediction issued by the model M is confirmed; otherwise (that is, if x is considered uncertain and hence $\pi(x) \in Bnd$), then the prediction issued by M is over-ruled causing it to (partially) abstain.[4]

In this article, based on the above mentioned categorization of applications of TWD in ML, we present a systematic survey of the specialized literature that emerged in the recent years since the proposal of TWD. Following previous observation summarized by Campagner et al. (2020a) , which noted how despite the increasing adoption of and interest toward TWD in ML, a lack of reporting standards and attention towards evaluation and reproducibility practices could be observed, we will discuss, in particular, a methodological analysis of the existing literature, to assess the reproducibility and reporting quality of existing studies. In particular, we will be interested in answering the following three questions:

1. What are the main characteristics of the TWD in ML literature from a scientometrics and analytical point of view. In particular: what is the nature of the studies concerned with TWD in ML (i.e., does the research tend mostly toward theoretical or empirical work)? Where are the main country hubs for research on TWD in ML?

[2] These issues have been widely studied in both RST, as well as in machine learning. In the former, the main questions of interest regard the notions of *indiscernibility* and *inconsistency* (Pawlak 1991; Pawlak and Skowron 2007) , while in the latter one the central notions of interest are that of a *decision boundary* and of *inductive bias*.

[3] One of the articles performed two sets of experiments, one in which cross-validation was applied, and one in which bootstrapping was applied.

[4] In this respect, we remark that significance evaluation in clustering studies (as opposed to classification ones) is particularly complex, due to the risk of *double dipping* (essentially, the act of using the same data to both perform clustering as well as perform statistical testing) which requires the application of selective inference approaches (Chen and Witten 2022; Gao et al. 2022) .

2. How does the current state-of-the-art of the research on TWD in ML fare with respect to reporting quality and study reproducibility?
3. What are the current trends for research on TWD in ML, according to the four above mentioned tasks, and which could be some particularly relevant future research directions emerging from the literature?

In order to answer these questions, the rest of this systematic review article will be structured as follows. In Sect. 2, we will describe the adopted reviewing methodology, as well as, in order to answer research question 2 above, delineate the criteria for the assessment of reporting quality of the surveyed studies. In Sect. 3, we will summarize the scientometrics and statistical results of our systematic review, in order to provide an answer to research question 1 above. In Sect. 4, we will summarize the findings of our results, with particular reference to our analysis of the reporting quality of the literature (so as to provide an answer to research question 2 above): based on these findings, in Sect. 4.5 we will provide clear indications for improvements as well as delineate potential directions for future research. Finally, in Sect. 5 we will summarize our contributions and provide some concluding remarks.

2 Methods

As mentioned in the Introduction, we conducted a systematic review of the literature regarding the application of TWD in ML, grounding on the above categorization of application in four different tasks: two related to uncertainty in the input of the ML process (i.e. handling of weakly supervised data, and handling of missing data), and two related to uncertainty quantification in the output of a ML model (i.e. uncertainty quantification in the output of classification algorithms, and uncertainty quantification in the output of clustering algorithms). To this aim, we surveyed the articles indexed by the Elsevier's Scopus database, applying four structured queries, as summarized in Table 1. We decided to focus exclusively on the Scopus database since previous research (Mongeon and Paul-Hus 2016; Thelwall and Sud 2022) showed it has more extensive coverage than other competing tools.

To discuss the collected articles, as well as for their analysis in terms of adherence to reporting and reproducibility standard, we considered a set of criteria extracted from the recent guideline proposed by Cabitza and Campagner (2021), especially those criteria that could be generalized to applications of ML outside the medical domain (for which that checklist was originally intended). In particular, we considered information related to three main semantic clusters, namely: general information, information about the experimental setting, and information about the model optimization and results.

In regard to the general information, we considered: authors' affiliation; the considered ML task; number of datasets considered in the experiments; application domains for the considered datasets; sources (including whether these were private or public sources) for the considered datasets; and datasets' dimensionality (i.e. number of features, instances, classes, etc).

In regard to the experimental setting we considered: type of validation (if any, internal validation, external validation, cross-validation, bootstrap, or variations thereof); evaluation metrics; information about significance testing and statistical analysis; summary of the main characteristics of the adopted TWD methodology; and type of output for the considered ML approach. Two main aspects of the experimental setting are particularly

Table 1 Queries for the Scopus database

Task	Query
Multi-rater, Multi-source, Multi-view and Weakly-Supervised (*weak supervision* task)	(TITLE-ABS-KEY ("three-way decision" OR "three way decision" OR "three-way decisions" OR "three way decisions") AND TITLE-ABS-KEY ("multi-rater" OR "multi-source" OR "multi source" OR "inter-rater" OR "inter rater" OR "interrater" OR "multiple raters" OR "multiple rater" OR "multi-rater" OR "inter observer" OR "inter-observer" OR "interobserver" OR "multiple source" OR "multiple sources" OR "multi-view" OR "multi view" OR "partial label" OR "semisupervised" OR "semi-supervised" OR "semi supervised" OR "weakly supervised" OR "weak supervision" OR "missing label" OR "superset" OR "partially labeled" OR "fuzzy label" OR "unlabeled") AND TITLE-ABS-KEY ("machine learning" OR "clustering" OR "feature selection" OR "feature reduction" OR "classification" OR "data analysis" OR "machine learning" OR "clustering" OR "feature selection" OR "feature reduction" OR "classification" OR "data analysis")) AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cp")) AND (LIMIT-TO (LANGUAGE , "English"))
Missing Data Management (*missing data* task)	(TITLE-ABS-KEY ("three-way decision" OR "three way decision" OR "three-way decisions" OR "three way decisions") AND TITLE-ABS-KEY ("missing data" OR "incomplete information" OR "data missing" OR "imputation" OR "missing" OR "incomplete" OR "interval data" OR "non-deterministic information") AND TITLE-ABS-KEY ("machine learning" OR "clustering" OR "data analysis" OR "feature selection" OR "feature reduction" OR "classification")) AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cp"))
Uncertainty Quantification in Classification (*classification* task)	(TITLE-ABS-KEY ("three-way decision" OR "three way decision" OR "three-way decisions" OR "three way decisions") AND TITLE-ABS-KEY ("machine learning" OR "data analysis" OR "learning") AND TITLE-ABS-KEY ("classification" OR "supervised learning" OR "supervised") AND NOT TITLE-ABS-KEY ("semisupervised" OR "semi supervised" OR "semi-supervised" OR "incomplete" OR "missing" OR "clustering" OR "multi-view" OR "multiview" OR "multi view" OR "multisource" OR "multi-source" OR "multi source" OR "partial" OR "partially" OR "co-training")) AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cp")) AND (LIMIT-TO (LANGUAGE , "English"))

Table 1 (continued)

Task	Query
Uncertainty Quantification in Clustering (*clustering* task)	(TITLE-ABS-KEY ("three-way decision" OR "three way decision" OR "three-way decisions" OR "three way decisions") AND TITLE-ABS-KEY ("interval clustering" OR "three-way clustering" OR "three way clustering" OR "rough clustering" OR "soft clustering" OR "interval-set clustering" OR "orthopartition") AND TITLE-ABS-KEY ("clustering" OR "unsupervised learning" OR "unsupervised" OR "cluster") AND NOT TITLE-ABS-KEY ("classification" OR "supervised" OR "active learning" OR "incomplete" OR "missing" OR "semisupervised" OR "semi supervised" OR "semi-supervised" OR "multi-view" OR "multi view" OR "multi source" OR "multi-source")) AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cp")) AND (LIMIT-TO (LANGUAGE , "English"))

remarkable and are thus worthy of further remarks. First, the type of validation: in this respect, we notice that external validation (in which the training and testing data come from two, not necessarily related, distributions) is considered the gold standard of evaluation practices for ML models, as it provides more reliable estimates than internal validation (in which the training and testing data come from the same distribution) since the latter can be subject to bias and overestimation of performance. A further distinction, however, should be made between different types of internal validation: “pure” internal validation (henceforth simply internal validation), in which training and testing are not clearly separated; hold-out validation, when the separation is determined by a single split chosen at random; k-fold, repeated and nested cross-validation, where multiple splits are considered by selection without replacement; bootstrap validation, which considers multiple splits by selection with replacement. Obviously, internal validation represents the least statistically sound form of validation, whose use is generally discouraged as compared to hold-out, cross-validation or boostrap procedures. Aside from the type of validation, a second important factor to consider when determining a result’s quality is whether the statistical significance of the result has been evaluated or not, to ensure that the outcome obtained is not due to chance. Statistical analysis, using either hypothesis testing or confidence interval analysis, is thus necessary to reduce the likelihood that the results are purely due to coincidence.

Finally, in regard to model optimization, we considered: information about missing data imputation; feature selection and hyper-parameter optimization (for any of them, whether it was performed, and using which methods); main hyper-parameters of the proposed methodology; and reported improvements according to the best comparison algorithm considered in the article. In regard to these aspects, the selected items were chosen to identify and report on the main factors influencing data leakage and estimation bias in ML studies, namely: imputation; feature selection; and hyper-parameter optimization. An imputation procedure replaces missing data with a single definite value. Since removing missing data can lead to a substantial decrease in the size of the dataset, these methods are used to allow to fully utilize a dataset without discarding potentially useful information. At the same time, if not performed carefully (e.g. by enforcing a strict separation between training

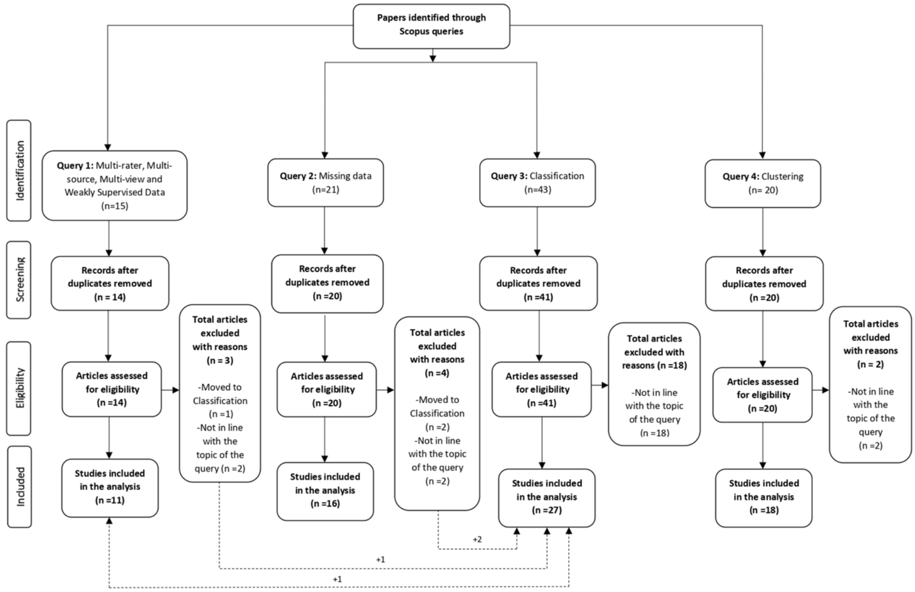

Fig. 2 The search procedure and the phase of study selection in the applied survey methodology

data and validation or testing data), imputation can lead to over-estimation of performance: thus, determining whether or not imputation was performed and in which manner within a given study can provide valuable insight as to its statistical validity. Feature selection, by contrast, refers to methods for selecting a subset of relevant attributes for their use in model development. Feature selection is of critical importance, because irrelevant or partially relevant attributes can negatively affect model performance by decreasing the accuracy of the model. Similarly to imputation, if not done correctly, feature selection can also lead to overfitting and overestimation of model performance due to data leakage. Finally, in regard to hyper-parameter optimization, we recall that hyper-parameters are parameters of a ML algorithm whose values are not directly estimated during the training process, but must instead set or selected a priori. Hyper-parameter optimization, then, is the process of choosing the optimal hyper-parameters of a learning algorithm so that it can optimally solve a machine learning problem. As with imputation and feature selection, this can help optimizing a Machine Learning model's performance, however, if not used carefully, it can lead to overfitting and data leaks.

Following the above mentioned criteria, we list the surveyed articles and their characteristics in the next sections. The search procedure and the phase of study selection are summarized in Fig. 2.

3 Results

In the following sections we review the queries and corresponding results. In particular, in Sect. 3.1 we report the results in regard to the weak supervision task; in Sect. 3.2 the results for the missing data task; in Sect. 3.3 the results for the classification task; and, finally, in Sect. 3.4, the results for the clustering task. Tables 2, 3, 4 summarize the results

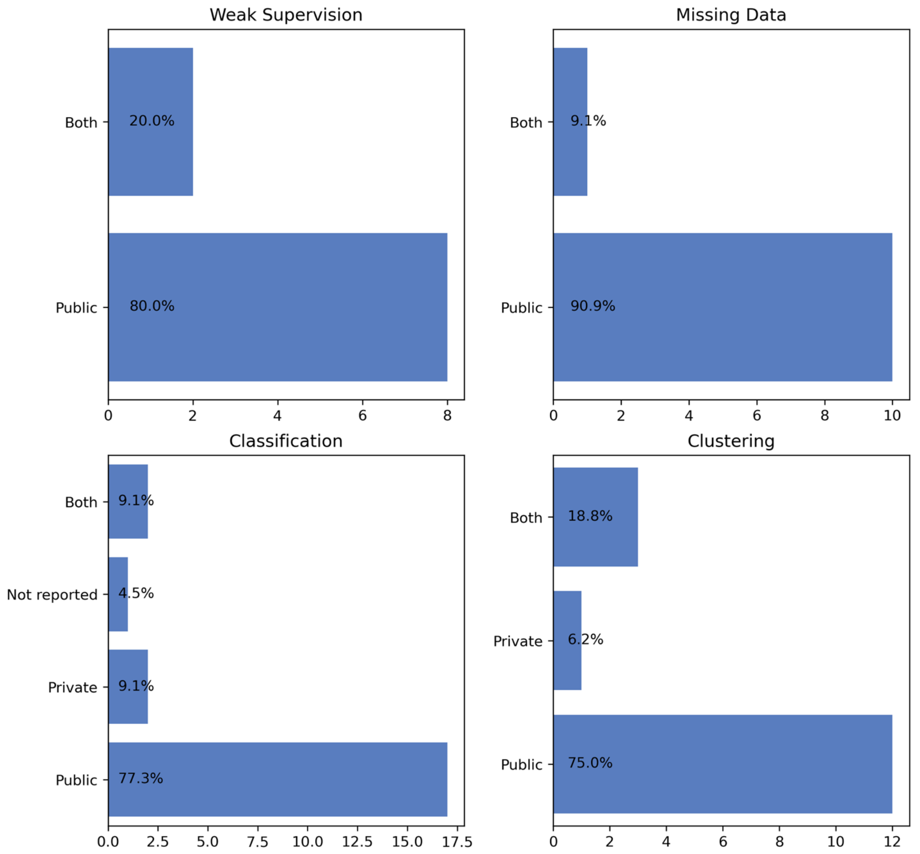

Fig. 3 Statistics about datasets' usage

for the weak supervision task; Tables 5, 6, 7 summarize the results for the missing data task; Tables 8, 9, 10 summarize the results for the classification task and, finally, Tables 11, 12, 13 summarize the results for the clustering task. All tables are in Appendix.

3.1 Weak supervision task

As a result of the query, 15 papers were returned. One of the articles was excluded because it was a duplicate, two other articles were excluded because they were not relevant to the query, and one article was moved to the classification task. In all, we included 11 studies whose collected data is presented in Tables 2, 3, 4.

More than half of the papers had authors who were affiliated with Chinese institutes (73%), followed by Italian (27%) institutes. Germany, Canada, and Poland were each represented by one article (9%). Ten out of eleven papers (91%) included an experimental section, while only one (9%) considered a theoretical analysis. Additionally, in the papers, the majority regarded classification tasks (73%), whereas clustering tasks were considered in 27%. In the reviewed papers, see Fig. 3, exclusively public datasets were utilized in 80%,

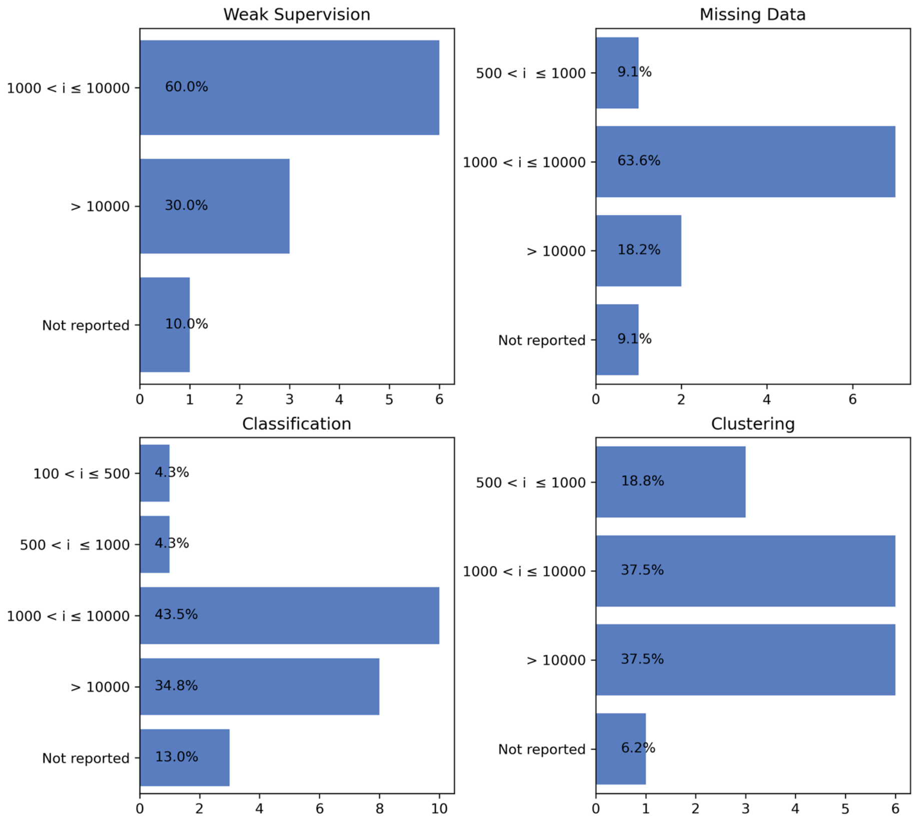

Fig. 4 Statistics about data dimensionality, in terms of number of instances

both private and public datasets were utilized in 20%, while no article considered only private datasets.

Excluding the unique theoretical paper, 60% of the articles reported having considered only datasets which had (at most) between 1000 and 10000 instances, 30% reported having considered also datasets with more than 10000 instances, while 10% reported no information on the number of instances (see Fig. 4). About 30% of the articles considered datasets with a number of features exceeding 100, 30% did not list any information concerning the features/attributes, while the remaining 40% of articles only considered datasets with less than 100 features (see Fig. 5). There were a majority of articles (50%) that did not state the number of classes for the used datasets, while the remaining 50% considered only binary tasks (20%) or tasks with less than 10 classes (30%) (see Fig. 6). Some of the articles provided additional types of information that we chose not to include in the diagrams and statistics (e.g. the number of raters, the number of views, the number of clusters).

In the validation experiments, internal validation was adopted more than 50% of the times, cross-validation approximately 27% of the times, and bootstrapping by approximately 9%, while one article (9%) did not report about the adopted validation method (see Fig. 7). On average, 70% of the experiments reported to have used accuracy

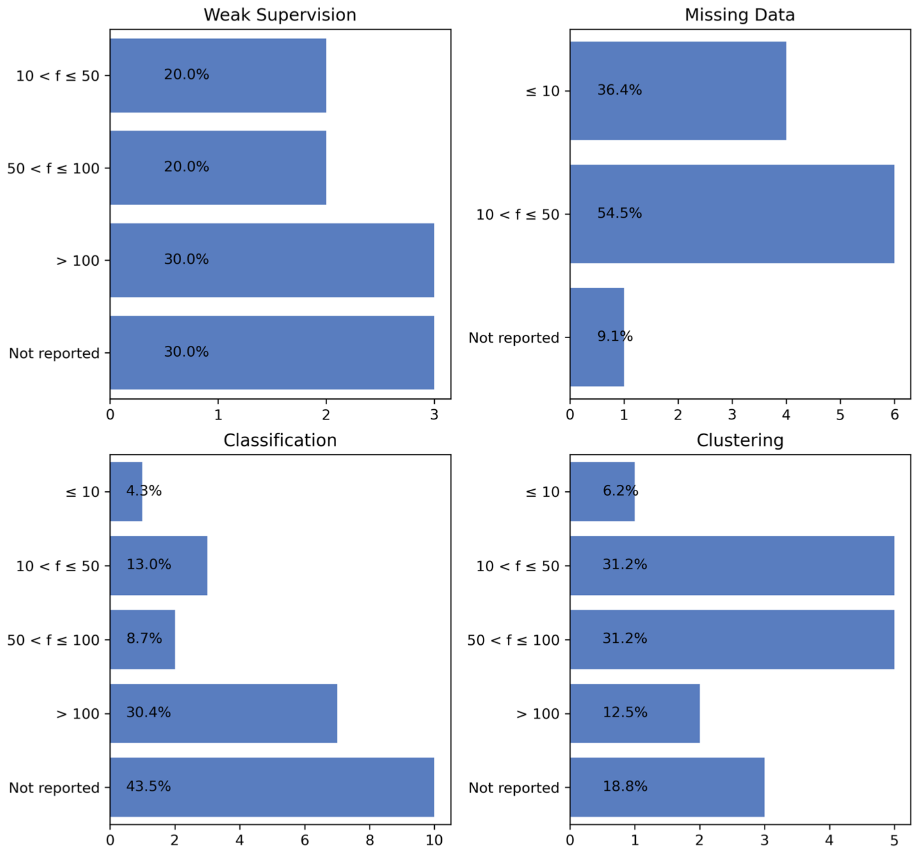

Fig. 5 Statistics about data dimensionality, in terms of number of features

as the evaluation metric, followed by 30% percent using NMI and 10% percent using Precision, while 40% reported using other metrics. Only 40% of the papers evaluated the statistical significance of the results (20% used confidence intervals). However, the majority of papers (60%) did not evaluate the significance of the results (see Fig. 8). In 90% of experimental designs, no imputation was performed, while 10% did not report whether or not it was performed (see Fig. 9). Nine articles (90%) did not perform feature selection, while one article (10%) reported having performed PCA (Principal Component Analysis), as shown in Fig. 10. Approximately 60% of experiments did not include any form of hyperparameter optimization, one article (10%) performed a parameter study, while 30% did not mention whether any form of optimization was performed or not (see Fig. 11).

3.2 Missing data task

As a result of the query, we obtained 21 records. One duplicate result was excluded, two other results were excluded from further research because they were not directly relevant to the query, and two results were moved to the classification task. Tables 5, 6, 7 show a summary of the collected data for each of the 16 included records.

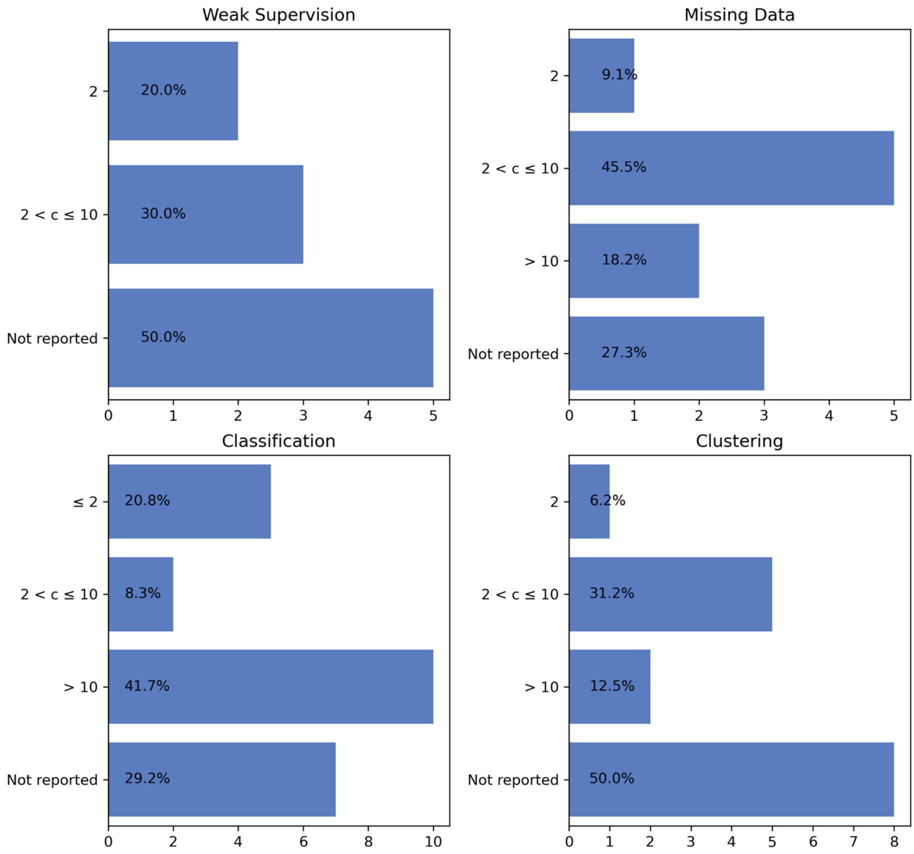

Fig. 6 Statistics about data dimensionality, in terms of number of classes

Researchers at Chinese institutions were represented in 75% of the surveyed articles, followed by Canadian (25%) and Japanese (25%) ones. Other affiliations counted for approximately 31% of the surveyed papers. 63% of the reviewed papers had an experimental section. In the papers, missing or incomplete data issues being the main focus of the study accounted for the majority (87%), while clustering and classification tasks were among the main aims in 44% and 31% respectively. In 91% of the studies, only public datasets were used, while only 9% combined public and private data (Fig. 3). 18% of the experiments reported having used datasets with more than 10000 instances, 64% datasets which had a number of instances between 1000 and 10000, 9% only dataset with less than 1000 instances, while 9% did not specify the number of instances (Fig. 4). The majority of datasets (55%) had features ranging from 10 to 50, while 9% didn't report enough information to determine the number of used features (Fig. 5). In most datasets (46%), there were between 2 and 10 classes, while 27% did not report the number of classes (Fig. 6). In 90% of experiments, only internal validation was performed (Fig. 7). Accuracy (90%) was the main evaluation metric, followed by F1-score (20%), coverage (20%) and other metrics (20%). A majority of the papers (80%) omitted to evaluate the statistical significance of the results. Only 20% of the papers evaluated the statistical significance of the results (of

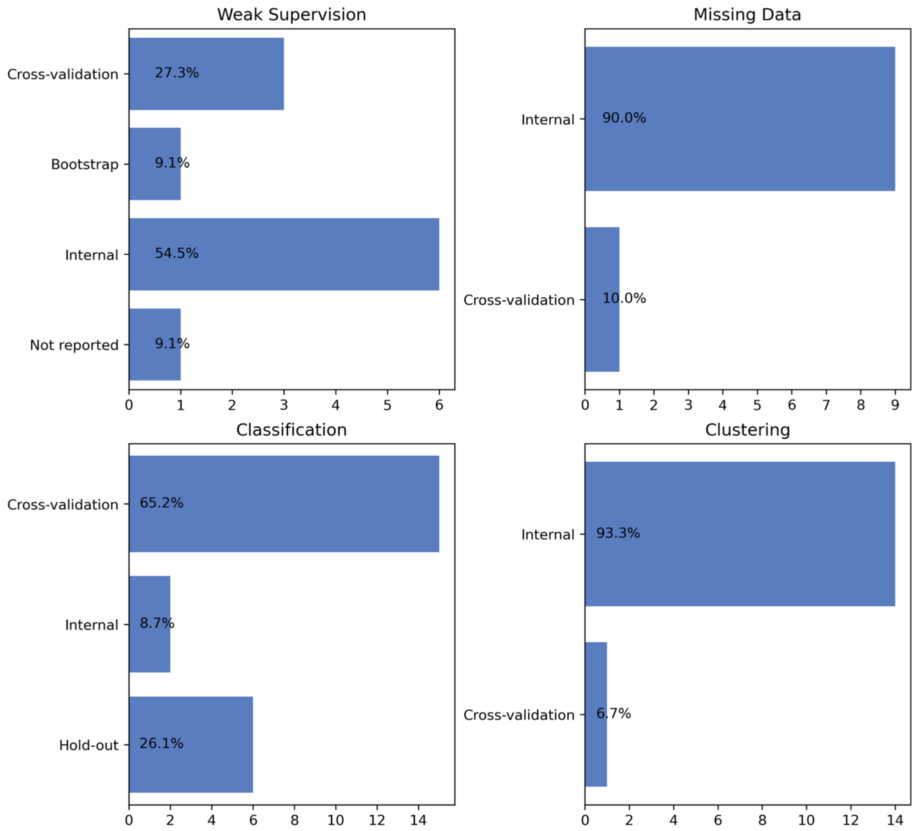

Fig. 7 Statistics about models' validation

which 10% used confidence intervals) (Fig. 8). In 60% of experiments, no imputation was performed, and 10% of experiments did not report whether imputation was performed. The imputation was performed in 30% of experiments (Fig. 9). In no study a selection of features was performed (Fig. 10). Approximately half of the experiments did not perform hyperparameter optimization, while the other half did not disclose whether it was carried out (Fig. 11).

3.3 Classification task

As a result of the query, we obtained 43 records. Two results were excluded from further research due to being duplicates, 18 results were further excluded because they were not directly relevant to the query. Three articles were included from queries 1 and 2. Thus, 27 articles were included in total, whose collected data is reported in Tables 8, 9, 10.

Researchers at Chinese institutions were represented in more than half of the surveyed articles (56%), followed by Canadian (22%) and Italian (15%) institutions. Almost all papers (85%) contained an experimental section after describing the content in a theoretical manner. In only 15% of the papers, the proposed three-way approach was

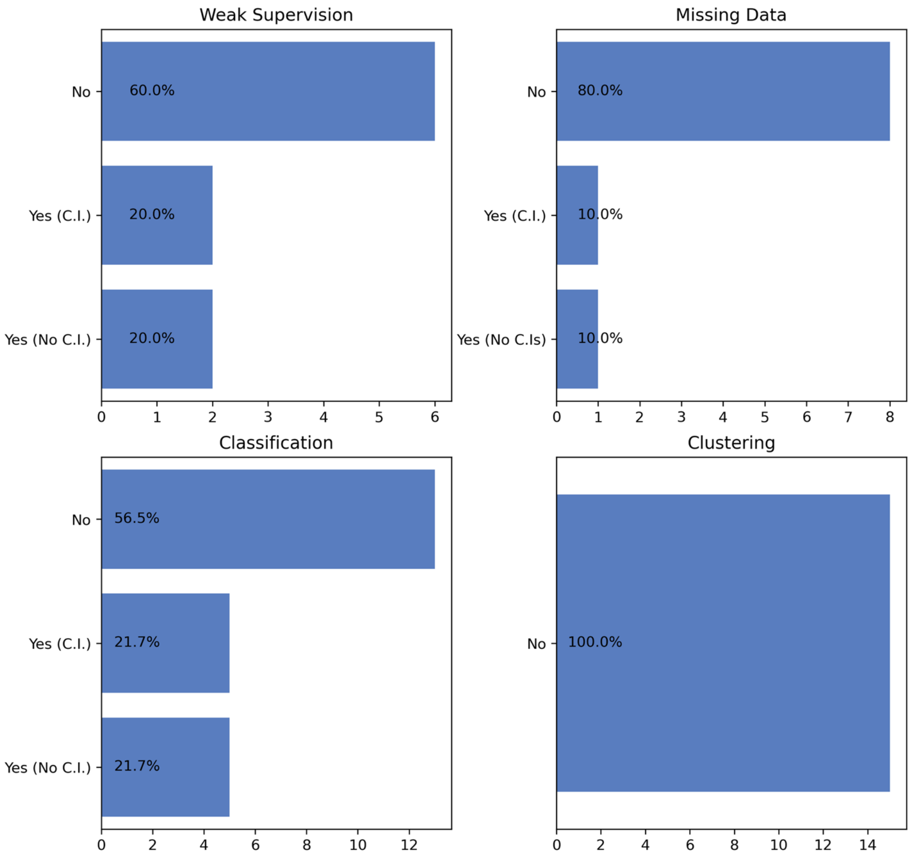

Fig. 8 Statistics about significance testing

not subjected to experimental validation, resulting in theoretical articles. The reviewed papers used exclusively public datasets in 74% of the articles, exclusively private datasets in 7%, 15% both, and 4% did not report the source of the datasets utilized (see Fig. 3). More than a third of datasets had more than 10000 instances (35%) or had between 1000 and 10000 (44%), while 13% of articles did not report the number of instances (see Fig. 4). Datasets with over 100 features were most common (30%), while 44% did not report the number of features/attributes (see Fig. 5). Most of the datasets (42% of them) considered more than 10 classes, while 29% did not mention how many classes they considered (see Fig. 6). As a validation method, cross-validation was adopted by 65% of experiments, hold-out validation by 26%, and internal validation by just less than 9% (see Fig.). In terms of evaluation metrics, accuracy was used the most (74%), followed by F1 (35%), Recall and Precision (30%), whereas other metrics covered 52% of the sample. A total of 43% of the papers analyzed the statistical significance of the results (22% used confidence intervals); the remaining 56% of them did not analyze the statistical significance of the results (Fig. 8). Moreover, imputation was not used by the vast majority of experiments (96%), while about 4.3% reported

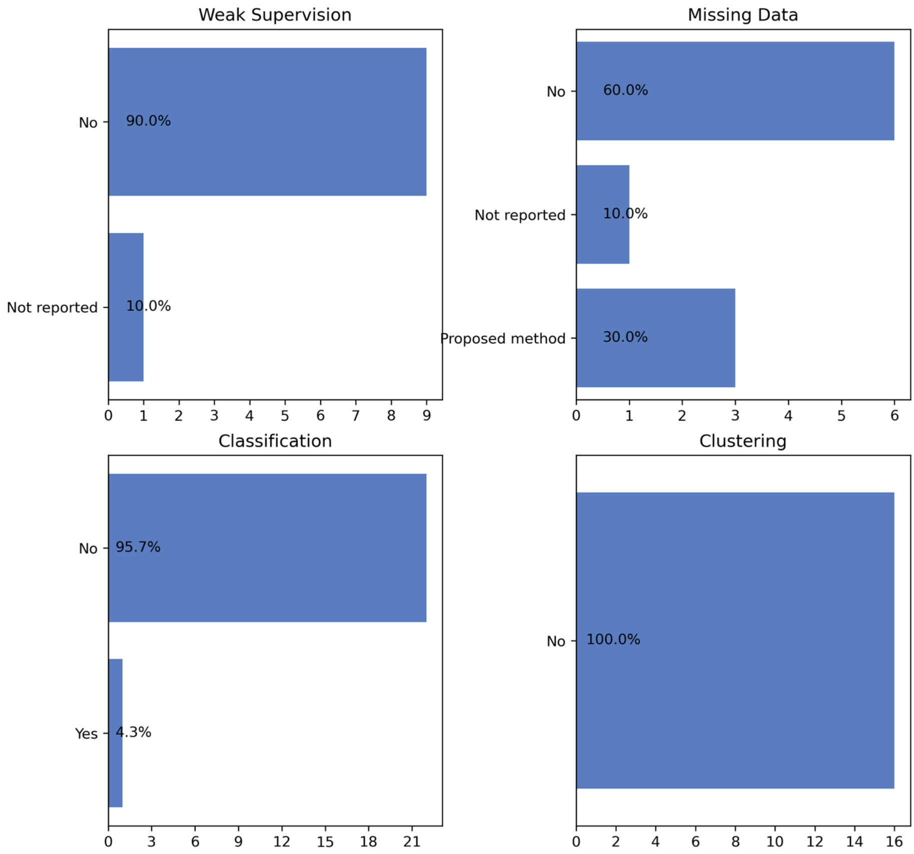

Fig. 9 Statistics about missing data imputation

using it (Fig. 9). In Fig. 10, regarding feature selection, 65% did not perform any feature selection, 22% of experiments used their proposed method, while the remaining 13% used other techniques. In regard to hyper-parameter optimization, 50% of the studies did not perform any form of optimization, in 27% of the studies a parameter study was performed, in 9% of the studies nested cross-validation was applied, while 9% of the studies either applied a new (not better defined) optimization procedure or did not report about hyper-parameter optimization (Fig. 11).

3.4 Clustering task

The query returned 20 records. Two of the results were excluded from further research since they were not directly relevant to the topic of the query. As a result, 18 studies were included, whose data is listed in Tables 11, 12, 13.

Almost all of the papers (94%) had at least one author affiliated with a Chinese institutions, while Canada was the second most represented affiliation (33%) A large majority of papers (83%) contained an experimental section after describing the content in a theoretical manner. The proposed three-way approach was not tested experimentally in 17% of the

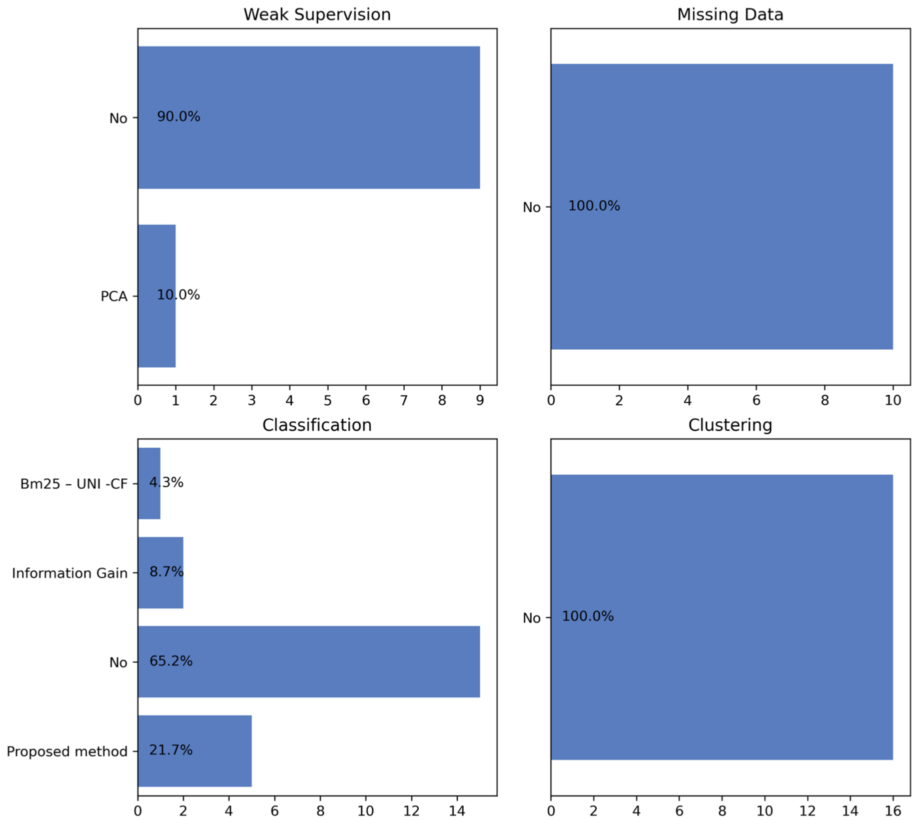

Fig. 10 Statistics about feature selection

papers, who only proposed theoretical methodologies. Approximately 75% of the studies used only public data, 6% used exclusively private data, and 19% used both private and public data (Fig. 3). Most of the studies used datasets with a number of instances greater than 10000 or between 1000 and 10000 (both 38% each). However, approximately 6% of articles did not report the number of instances (Fig. 4). Most of the studies considered datasets with between 10 and 50 or between 50 and 100 features (31% each), whereas 19% of papers did not state the number of features (Fig. 5). The number of classes was not provided in half of the studies (50%), while in the other half, the number most frequently ranged between 2 and 10 (31%), or more than 10 (13%). The remaining 6% reported using datasets with only 2 classes (Fig. 6).

There were 93% of experiments which performed validation by using internal validation, while approximately 7% used cross-validation (Fig. 7). In the studies, accuracy (81%) was the most frequently used evaluation metric, followed by two internal quality metrics, i.e. Davies–Bouldin (31%) and Silhouette (31%) indices, and the external quality metric NMI (25%). Even though the proposed methods led to some improvements for all of the considered papers, no statistical significance procedure was applied in any of the experiments (Fig. 8). In all the experiments, no imputation was performed on the data

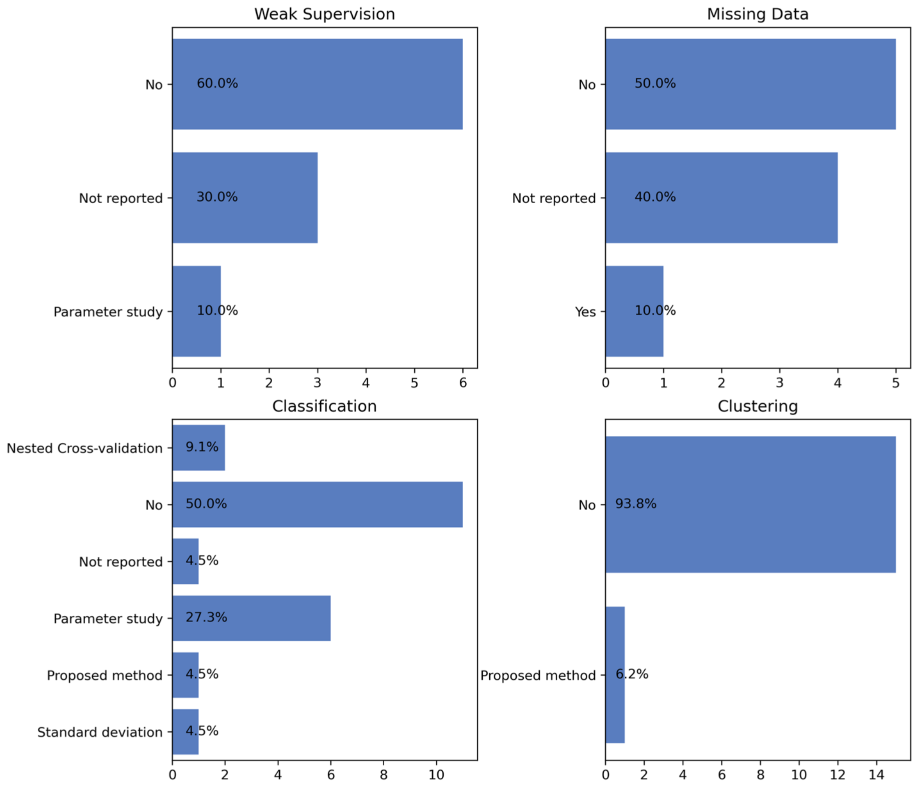

Fig. 11 Statistics about hyper-parameter optimization

(Fig. 9). None of the experiments employed feature selection in any way (Fig. 10). All but one of the experiments (94%) did not involve any form of hyper-parameter optimization (Fig. 11).

4 Discussion

In this section we discuss the main findings emerging from our systematic analysis of the literature, focusing firstly on the broader observations shared among the four considered ML tasks. In general, we observed that the main hubs for research on TWD in ML, for all tasks, were associated with Chinese affiliations, followed by Italian and Canadian ones, and then Japanese and Polish ones: such a picture is not particularly surprising, since a large portion of researchers associated with TWD and RST, including some of those who made seminal contributions to TWD research (Yao 2012; Yu 2017) , are affiliated with institutions in these countries. Furthermore, almost all of the surveyed articles largely focused on the experimental evaluation of proposed algorithmic approaches, rather than on theoretical contributions: this finding reflects an analogous trend in the ML literature, where, since the advent of deep learning, research has focused more on the engineering and experimental aspects of the discipline, rather than on the theoretical ones (Pugliese et al. 2021) . Interestingly, in the ML literature, there have recently been calls for a more balanced approach aimed at bridging theory and practice as a way to enable deeper understanding about the functioning of modern ML models, as well as to provide actionable and rigorous advice on how to select ML solutions for particular application (Hutson 2018) : in this light, and given the above mentioned trend in the TWD literature, the more systematic exploration of the theoretical aspects of TWD, and how they interact with ML theory, could be a research direction of general interest.

In the following sections, we will explore in greater detail the insights and observations relative to the four ML tasks we considered in this survey, as well as, in Sect. 4.5, discuss implications for research and future research directions that can be drawn from our analysis of the literature.

4.1 Weak supervision task

We start the discussion of the reported results from the weak supervision task. In this regard, we remark that approximately 30% of the studies surveyed in the *weak supervision* task did not report any information regarding data dimensionality, in terms of either number of instances or features. This is an aspect that may limit the reproducibility of studies (McDermott et al. 2021) . Indeed, for public datasets, it is unknown whether all of the original dataset contents were utilized or merely a part of them. Even more significantly, we note that the majority of studies did not provide any indication on the number of classes, leaving it unknown whether all classes or merely sub-tasks have been considered. These observations highlight a lack of adoption of reporting and reproducibility standards (Boyd 2021) for applications of TWD to weak supervision tasks.

Remarkably, only a minority of articles in the weak supervision task encompassed a validation based on cross-validation or bootstrap, with the majority of surveyed papers only applying an internal validation. This latter finding may have severe consequences as it can be a cause of data leakage (Bussola et al. 2019) , undermining the reliability of these studies and raising the risk of overly optimistic performance estimates. Furthermore, these previously mentioned issues may in turn lead to problems with generalizability as the use of internal validation limits the applicability of the reported improvements to other settings (Steyerberg and Harrell 2016) . In this regard, no study considered an external (or internal-external) validation: while this is not a problem per se, it makes evaluating the robustness of the proposed methods to data or concept shifts, or similar distributional issues, more difficult (Cabitza et al. 2021) .

As a further problem, in the weak supervision task, a significant number of articles did not report about the execution of hyper-parameter optimization (around 30%): this may significantly affect the evaluation of the generalizability of the respective studies, especially in light of the fact, as mentioned above, that most studies (and, in particular, all of those that did not report about performing hyper-parameter optimization or not) only performed internal validation. The above mentioned issues may lead even to reproducibility problems (Dodge et al. 2019) , as the hyper-parameters optimization stage would require additional information regarding the assumptions made.

In regard to the evaluation of the proposed approaches, most of the surveyed articles only considered accuracy, without any information about other metrics such as sensitivity or specificity (or the Area under the ROC curve, AUC). Especially in light of the fact that none of the studies reported whether the considered datasets were imbalanced or not (Japkowicz 2013) , this gap may result in a risk of performance overestimation, and it does not allow to understand the error patterns of the proposed methods (i.e. whether they favor false positives or false negatives). Moreover, only a minority of the studies applied some procedure for significance analysis, making the reported improvement in performance w.r.t. to the state of the art dubious (Benavoli et al. 2016; Demšar 2006) .

In regard to the methodologies considered for the implementation of TWD in the *weak supervision* task, interestingly, approaches based on three-way clustering were widely represented among the surveyed papers, with 50% of the papers applying some kind of clustering-based approach (including label propagation (Xiaojin and Zoubin 2002)). While this finding is not per se surprising (indeed, clustering-based approaches are used also in many approaches not based on TWD (Afyouni et al. 2022; Chao et al. 2017; Shao et al. 2015; Zhou 2018)), we note that the applicability of the assumptions which are typically required for clustering-based approaches (namely, manifold regularity or Lipschitz-ness assumptions) for this task have been criticized in the specialized literature, due to difficulties in ensuring proper generalization in high-dimensional contexts (Assent 2012) . Remarkably, in this sense, two of the studies which adopted a clustering-based approach were evaluated also on very high-dimensional datasets and reported good results. This results may suggest that the application of TWD-based clustering techniques (rather than standard hard clustering ones) could provide some advantages in regard to robustness to the curse of dimensionality. Nonetheless, due also to the above mentioned issues in regard to reproducibility and generalizability of the results reported in the surveyed study, further work should be devoted at investigating this purported advantage of TWD.

4.2 Missing data task

As for the weak supervision task, also for the missing data task a large part of the surveyed studies lacked information regarding instances, features, and most significantly, classes. Similarly, most studies only performed internal validation and did not report about whether hyper-parameter optimization was performed or not, undermining the reliability of the reported results.

Interestingly, despite the management of missing data being the main focus of the missing data task, 60% of the paper did not involve any form of imputation. We believe this latter observation to be a particularly remarkable as it highlights the fact that the application of TWD inspired approaches allows to handle missing data without performing any kind of missing value replacement. By contrast, missing value replacement is the most popular way (along with missing indicators) to handle this type of data in standard ML pipelines (Lenz et al. 2022) : as imputation is one of the main sources of data leakage and overestimation of performance (Kapoor and Narayanan 2022) , if not performed carefully , TWD-based approaches for missing data management might then offer some benefits for reproducibility and generalizability, since we observed how they generally avoid imputation .

However, similarly to what we previously reported for the weak supervision task, in the missing data task the problem with metrics persists, even though in this case two papers reported about the F1-score (which is a more balanced account on performance than accuracy) and one of them reported also the precision and recall. Furthermore, the problem of lack of statistical significance is even more pervasive than for the *weak supervision* category: 80% of the studies did not perform any procedure for the assessment of statistical significance. Nonetheless, in the specific situation of the missing data task, our review reveals how TWD implementations in ML, despite the limited generalizability and repeatability of the studies, eliminate one of the key factors contributing to the overestimation of performance, which is represented by imputation.

4.3 Classification task

With regard to the classification category, the problem of scarce reporting about the number of instance, features or classes is even larger than in the two previous tasks, with around 40% of the studies which did not report about either the number of features or the number of classes. At the same time, the validation practices adopted by studies concerned with classification approaches were more robust, with most studies adopting some form of cross-validation or hold-out validation (around 80%) and only a minority of them (under 10%) adopting only internal validation.

In contrast to the two previous tasks, a larger number of works reported some balanced performance metrics along with accuracy (around one third), however in most cases only accuracy was reported as a measure of error rate. Remarkably, even though the classification task regarded the use of three-way decision as a way to implement uncertainty quantification (Hüllermeier and Waegeman 2021; Kompa et al. 2021) through either rejection (Hendrickx et al. 2021) or partial abstention (Mortier et al. 2021) , only a small minority of studies reported some measure of coverage or efficiency (4 out of 25). Such a lack of information makes the evaluation of the reported performances hard to analyze and assess, as the reported improvements could be caused largely by a small coverage of the proposed methods (Nadeem et al. 2009) . Indeed, TWD-based ML approaches, similarly to other cautious inference methods, aim to strike a trade-off between reduced coverage and higher accuracy (Golfarelli et al. 1997; Lars Kai et al. 1997; Nadeem et al. 2009) : without any information on the first component of this trade-off, however, it is impossible to evaluate whether such methods did really provide any kind of benefit compared to state-of-the-art methods. This is a very critical point considering the scope of this review, which is centered on applications of TWD in ML.

Despite this latter issue, however, compared to the previous two tasks, a larger number of studies applied some kind of procedure for significance testing, with just less than one half of the surveyed articles using either hypothesis testing or confidence intervals. Together with the fact that most studies correctly reported about the application of either feature selection and hyper-parameter optimization, these observations make the findings reported in studies concerned with the classification task the most reliable ones from a statistical point of view, even though some improvements (in particular in regard to the application of significance testing) should still be achieved.

Among the adopted techniques to implement TWD-based uncertainty quantification in the output, the most frequently represented ones were Decision Theoretic Rough Sets or other Rough Set-based models. This finding is not particularly surprising, as TWD originally emerged from the study of Rough Set theory (Yao 2010) . At the same time, this finding explains the large percentage of studies in which both classification and feature selection approaches were proposed, since Rough Set theory can be applied to both these tasks (through reduct search and rule induction, respectively) (Bello and Falcon 2017; Pawlak 1991) . Notably, however, this last observation highlights the need to conduct ablation studies (Lipton and Steinhardt 2018) , which were not performed in any of the considered studies. Ablation studies are of fundamental importance to understand whether the reported improvements are solely or largely due to only one part of the proposed approach (e.g. only to the feature selection component), and to decompose the contribution of each of the respective components.

4.4 Clustering task

Clustering was the task in which there were a larger portion of studies which did not perform any kind of validation or assessment, as well as the one in which the larger number of private datasets were used (more than double, and almost triple, the percentage than that of other tasks). As mentioned previously, this may impact on the reproducibility of the reported results, which may only be partial and restricted to the public datasets. This is especially relevant since more than 1 out of 4 studies used at least one private dataset. The problem of lack of reporting about the number of classes (or clusters), instances and features was less relevant for the clustering task than for the other ones: only 2 of the considered studies did not report about this information, chiefly due to the fact that the considered datasets were not labeled.

By contrast, almost all of the studies performed only an internal validation. However, this is a much less critical problem for clustering than for other tasks, since clustering is usually adopted in a transductive fashion as a way to perform knowledge discovery (Trivedi et al. 2015) . At the same time, the problem of significance of the results is particularly critical, as none of the considered studies reported having applied any such procedure. This makes it impossible to evaluate the statistical soundness and reliability of the reported results, especially in light of the observations above. However, none of the considered studies applied any form of imputation (because datasets were complete), feature selection or hyper-parameter optimization, making the risk of data leakage marginal as compared to other tasks.

In regard to metrics, most studies applied both internal and external validation criteria (Rendón et al. 2011) : in particular, none of the studies applied only internal criteria, whose utility as measures for objective clustering evaluation has been questioned in recent studies (Arbelaitz et al. 2013; Lei et al. 2017; Ullmann et al. 2022) . Nonetheless, it is to note that all of the studies did apply only performance measures for hard clustering algorithms (Denoeux et al. 2017) . Indeed, none of the considered studies applied performance measures that allow to take into account the amount of objects placed in the boundaries of some cluster (Campagner and Ciucci 2019) or, more in general, to quantify the uncertainty and ambiguity in the output of the corresponding algorithm (Campagner et al. 2023a, b) . This is a rather relevant problem, which was already reported in our previous review (Campagner et al. 2020a) and other previous contributions (Denoeux et al. 2017; Hullermeier et al. 2011) , for two main reasons. On the one hand, such an evaluation does not allow a fair comparison between algorithms which belong to different algorithmic families (Campagner et al. 2023a, b) and, especially so, between TWD-based and hard clustering algorithms (Campagner and Ciucci 2019) , as these two types of algorithms feature a completely different type of output. On the other hand, because of a lack of clarity about how boundary objects are to be treated (are they considered as erroneous in regard to cluster placement? or as being correct assignments?) and, more generally, about the semantics assigned to these objects (Campagner et al. 2023a, b) (are they intended to represent some form of uncertainty? or rather some degree of overlap among clusters?). Thus, we remark that evaluation results reported in the surveyed articles, and especially so in regard to those results that compare hard clustering and TWD-based clustering methods, may be highly biased, as similarly reported in regard to the *classification* task. We believe, thus, that more attention should be devoted at the investigation of TWD-based clustering approaches through appropriate evaluation metrics.

Interestingly, compared to other tasks, there was a much higher variety of proposed methodologies, with no particular approach being significantly more represented than others. However, a relevant number of methods were based either on partitional clustering (e.g. algorithms in the three-way k-means family (Yu 2017)) or density-based clustering. Interestingly, almost all studies considered approaches based on the three-way clustering formalism, formulated by Yu (2017), rather than other competing formalisms for implementing three-way decision in clustering, e.g. rough clustering (Lingras and West 2004) or interval-set clustering (Yao et al. 2009) . Compared with these latter formalisms, three-way clustering allows to distinguish more clearly between two types of uncertain objects (Campagner et al. 2022) , i.e. between-cluster objects (objects that are placed in the boundary of at least two clusters) and outlier-like objects (objects that are placed in the boundary of only a single cluster). While this is an interesting property of three-way clustering for the purpose of uncertainty quantification, we remark here that none of the studies did evaluate differences in the considered algorithms in regard to possible trade-offs between these two forms of uncertainty since, as discussed above, studies only reported measures for hard clustering, largely disregarding objects in the boundaries.

4.5 Implications for research and future directions

In the previous sections, we discussed our main results concerning the analysis of the TWD literature in reference to the four considered ML tasks, identifying weak areas as well as suggesting potential areas for improvement and directions for future research. In this section, we summarize the main general indications that could be helpful to researchers in TWD and its applications in ML.

A first indication emerges from the observed lack of reporting standard, both in regard to data and model aspects:

- A majority of studies failed to comprehensively document the main characteristics of the datasets considered for validation, including such basic information as the number of considered classes. Such lack of information can have a severe impact on the reproducibility, and hence credibility, of a study's results, especially when compounded with the use of private datasets (as in the clustering category) where reproducibility is impossible, by definition. A possible solution to this problem would be for future studies in the TWD literature to adopt and follow reporting checklists, including both checklists devoted specifically to data aspects (Boyd 2021) as well as more general reporting guidelines;
- Many studies (especially so in the weak supervision and missing data categories) also failed to provide sufficient details on crucial aspects of the data science pipeline, including information about hyper-parameter optimization and related tasks (e.g., feature selection) which could severely impact on the generalizability and robustness of the reported results. As for the previous point, a possible solution to improve the reporting standard in the TWD literature would be for future studies to more closely follow existing standard reporting guidelines, such as the one we adopted in this article (Cabitza and Campagner 2021) or related ones (Crossnohere et al. 2022) . While most of these checklists have originally been proposed in the context of medical applications (where, indeed, the need for standards that ensure reporting quality and reproducibility is particularly critical), general principles can be easily drawn from them.

Another weak point in the surveyed studies concerned the validation of the proposed methodologies. In this respect, both the adopted validation designs, the selection of validation metrics, as well as the statistical analysis of results were found to be lacking:

- In regard to validation design, a surprisingly large number of studies only adopted internal validation study designs, where training and testing of a ML approach are performed on the same set of data. While these validation designs are not wrong by themselves (indeed, much of theoretical ML research is devoted to exploring what can be said about the generalizability of ML methods using only internal validation (Shalev-Shwartz and Ben-David 2014)), their application may limit the generalizability and trustworthiness of the results, if the risks of overfitting and data leakage are not properly accounted for. The simplest solution to this problem would consists in exclusively adopting validation designs that enforce a strict separation of training and testing data, such as hold-out validation or cross-validation, or also designs that employ randomization to correct for the risk of overfitting, e.g. bootstrapping. These validation designs are by now commonplace in the ML literature, hence, it was surprising they were not extensively adopted in the TWD literature (with the exception of the classification task). Notably, however, we note that also these validation designs are not sufficient to prove the generalizability of ML techniques in out-of-distribution or related settings (Cabitza et al. 2021; Steyerberg and Harrell 2016) . With this respect, it is relevant to remark that none of the considered studies employed external validation or related designs: thus, we believe that exploring the robustness of TWD-inspired methods in these settings could be an interesting direction for future research;
- In regard to the adopted validation metrics, most studies (especially so in the weak supervision and missing data tasks) only focused on accuracy. Despite being widely used, accuracy is not well suited for settings affected by label imbalance, where the entire confusion matrix (and derived metrics, such as sensitivity, specificity, positive and negative predictive values) can be more informative. Furthermore, almost none of the surveyed studies considered metrics that go beyond the measurement of discrimination power (i.e., error rate), neglecting important performance dimensions such as calibration (Francisco M et al. 2023) . As with the previously noted issues related to reporting quality, also in this respect the adoption of reporting guidelines could help TWD researchers in the selection of appropriate validation metrics and related tools (e.g., visualizations);
- Finally, in regard to statistical analysis, only a minority of studies assessed the significance of the observed results. Statistical analysis is important to provide solid evidence concerning the studies' results and derived conclusions (Demšar 2006) , especially when the objective of such a study is to prove that a proposed TWD-based method provides better performance than the state-of-the-art. To this end, it is recommended that future work in TWD research provide more comprehensive statistical analysis of the reported results, adopting approaches either based on hypothesis testing (Demšar 2006) or confidence intervals (Berrar 2017) : importantly, following recent guidelines on the subject, TWD researchers should not only report about the significance of results, but instead focus on providing comprehensive discussion of p-values, effect sizes (Greenland et al. 2016) as well as potential corrections needed to avoid biases and over-estimation of effects, e.g. correction for multiple hypothesis testing (García-Pérez 2023) .

Concluding, we also provide potential suggestions for future directions of research in the application of TWD to ML, as emerged from our results:

- Clustering-based TWD approaches for weakly supervised learning seem to improve robustness to the curse of dimensionality, especially in comparison with traditional clustering-based methods. This hypothesis should be further investigated in future research;
- TWD-based approaches for missing data management seem to offer a distinct advantage over traditional techniques adopted in the ML literature, in that they do not necessarily require (and usually do not use) imputation, a data processing step that may negatively impact the generalizability of ML studies. Future research should be devoted at exploring this advantage of TWD-based methods, as well as at comparing them with other ML techniques that likewise do not require imputation (e.g., missing indicators);
- In regard to the classification task, we noted how original TWD-based approaches typically combined a feature selection step with a classification one: this characteristic derives from the widespread usage of techniques inspired by rough set-theoretic methods in the TWD literature. On the one hand, this should drive the literature to conduct ablation studies aimed at decoupling the impact on performance of the feature selection and classification components: we believe such studies could be especially relevant for identifying particularly effective feature selection methods, as well as ways to mix and match different components in a more systematic way (e.g., by employing hyper-parameter optimization procedure). On the other hand, the extensive focus on techniques inspired by RST leaves open the possibility to explore other TWD-based methodologies that do not rely on such approaches, with particular reference to synergistic approaches that combine TWD with other cautious inference or related approaches, such as conformal prediction or active learning;
- In regard to both the classification and clustering tasks, we observed that only a minority of the surveyed articles properly accounted for the uncertainty quantification properties of TWD-based approaches. As for the classification task, we believe that future research should be focused at better exploring the accuracy-coverage trade-off offered by commonly adopted TWD-based methods, both from an empirical point of view (indeed, as we noted, only few works reported the coverage of the proposed methods) as well as from a theoretical one. As for the clustering task, we believe that future studies that more accurately and precisely investigate their advantages with respect to hard clustering methods are particularly needed;
- Finally, in regard to the clustering task, we noted how most of the surveyed studies focused on techniques based on generalizing existing partitional (e.g., k-means) and density (e.g., DBSCAN) clustering methods to the framework of three-way clustering. On the one hand, this suggests that further attention should be focused toward other clustering methods' families, such as hierarchical clustering, which may be better suited for specific applications. On the other hand, due to three-way clustering's ability to more comprehensively represent clustering uncertainty (w.r.t. to rough clustering and interval set clustering), a particularly interesting direction for future research would be the investigation of the trade-offs between these different forms of uncertainty.

5 Conclusions

In this paper, we comprehensively surveyed and assessed the main contributions regarding TWD in the specialized literature. This extends and complements a recent narrative review (Campagner et al. 2020a) , which offered a taxonomy of applications of TWD-based approaches in ML grounding on the distinction between strategies that deal with uncertainty either in the input or output of a ML pipeline. We adopted the above taxonomy and performed a systematic review focusing on four tasks: learning from weak supervision and missing data management, in regard to the application of TWD that handles uncertainty in the input, and uncertainty quantification in classification and clustering for those regarding uncertainty in the output.

In general, despite the increasing popularity of TWD and the increasing number of related successful studies (even in comparison to more traditional ML approaches), we highlighted that the sound application of evaluation best practices and adoption of reporting standards are still a rare occurrence. For this reason, we provided clear indications for improvement in reporting and reproducibility, which we believe could be useful to improve the methodological and conceptual contributions that TWD approaches may offer to the ML community and scholarly discipline. Moreover, through our review we highlighted some particularly relevant advantages and peculiarities offered by TWD, which we believe could be object of, or motivate, future research:

- Under the *weak supervision* category, we highlighted how the implementation of TWD-based clustering techniques (as opposed to hard clustering ones) could bring some benefits in terms of robustness to the curse of dimensionality. These results should be further explored and validated;
- In the discussion of the *missing data* task, we remarked how the application of TWD inspired approaches enables the management of missing data without performing any form of imputation. Since imputation is one of the primary sources of data leakage and overestimation of performance (Kapoor and Narayanan 2022) , if not handled appropriately, this line of inquiry may bring some benefits for the reproducibility and generalizability of ML studies involving missing data management steps , as TWD approaches may limit this source of bias ;
- In our review, in regard to the handling of uncertainty in the output of the ML pipeline, we focused mainly on the classification and clustering task. However, also other tasks exist for which uncertainty quantification can be applied, such as regression or forecasting. However, such tasks have scarcely been considered in the TWD literature: future work should thus be devoted to the investigation of applications of TWD to these tasks;
- Furthermore, our review revealed that, particularly for output-related tasks, most studies have so far neglected to address relevant assessment metrics in regard to the type of output (classificatory or clustering) under consideration. We believe that this feature, if appropriately pursued, would unleash the full potential of TWD techniques in the field of ML to be realized;
- Finally, we believe that the possibility to provide partial abstentions as a form of output could enable the investigation of TWD techniques in the field of human-machine interaction, as a way to mitigate the risk of emergence of automation-related biases, as well as its study in relation with close sub-fields of ML, such as active learning or machine teaching.

Appendix: results of the queries

In this Section, we report the results of the queries, in tabular form

Table 2 Results for the weak supervision task: general information

Paper	Authors' affiliations	Task	Number of datasets	Datasets' domains	Datasets' sources	Dimensionality
Campagner et al. (2020b)	Italy	Partial labels Classification	6	Medicine Synthetic General	IRCCS galeazzi (private) UCI	Instances: 150 to 10^6 Features: 4 to 64 Classes: 2 to 5
Campagner and Ciucci (2018)	Italy	Semi-supervised Classification	Theoretical Paper	N/A	N/A	N/A
Campagner et al. (2021b)	Italy Germany	Multi-rater Classification	6	Medicine Synthetic	Private UCI	Instances: 617 to 10000 Features: 8 to 2500 Raters: 11 to 21
Chen et al. (2015)	China	Multi-view Supervised Classification	2	Text (general) General	UCI	Instances: 3196 to 4601 Features: 36 to 58 Classes: 2
Dai et al. (2019)	China	Semi-supervised Classification	2	Images (faces)	Ohio State Univ.[8] AT &T Labs[5]	Instances: 1600 to 1764 Classes: 2
Shengdan et al. (2022)	China Canada Poland	Semi-supervised Active Leaning Classification	8	Medicine Images General	UCI	Instances: 178 to 2310 Features: 7 to 41 Classes: 2 to 10
Chenchen et al. (2017)	China	Multi-source Classification Concept Learning	5	Medicine Images(general)	UCI	Instances: 20000 to 1025010 Features: 1 to 20 Classes: 3 to 10
Xiong and Yu (2019)	China	Multi-view Clustering	4	Generalg Text (news) Images (digits) Sensors (vehicles)	UCI 3Sources[2] SensIT[6]	Instances: 169 to 2000 Features: 6 to 3631 Views: 3 to 10
Yu et al. (2017b)	China	Multi-view Semi-supervised Clustering	4	Text (general) Scientific Sensors (vehicles)	UCI Citeseer[5] Cora[5] SensIT[6]	Instances: 300 to 3302 Views: 2 to 3

Table 2 (continued)

Paper	Authors' affiliations	Task	Number of datasets	Datasets' domains	Datasets' sources	Dimensionality
Yu et al. (2020b)	China	Multi-view Clustering	7	Text (news) Text (webpages) Scientific Sensors (vehicles) Movies Images(digits)	UCI 3Sources[2] IMDb[6] WebKB[5] Citeseer[5] Cora[5] SensIT[7]	Instances: 169 to 3312 Features: 50 to 3703 Views: 2 to 4
Zhu et al. (2020)	China	Multi-view Multi-label Semi-supervised Clustering	4	Text (news) Images (general)	VOC[8] MIR-Flickr[9] 3Sources[10] NUS-WIDE[11]	Not reported

Table 3 Results for the weak supervision task: experimental setting

Paper	Validation type	Evaluation metrics	Significance testing	Proposed TW methodology	Output type
Campagner et al. (2020b)	Cross-validation	Accuracy	Friedman Test	Pseudo-label Belief function Interval Optimization	Single label
Campagner and Ciucci (2018)	N/A	N/A	N/A	Decision Tree	Single label with abstention
Campagner et al. (2021b)	Cross-validation Bootstrap	Accuracy	Confidence Intervals; Friedman Test	Ensemble learning; Pseudo-labels	Single label
Chen et al. (2015)	Cross-validation	Accuracy	No	Three-way Classification	Single label
Dai et al. (2019)	Internal	Cost	No	Co-training Sequential	Single label with abstention
Shengdan et al. (2022)	Not reported	Accuracy	Confidence Interval	Active Learning Label propagation	Single label
Chenchen et al. (2017)	Internal	Execution Time	No	Three-way Concept Learning	Three-way Cognitive Concepts
Xiong and Yu (2019)	Internal	NMI Accuracy ARI	No	Three-way Clustering Spectral Clustering	Three-way Clustering
Yu et al. (2017b)	Internal	Accuracy NMI	No	Three-way Clustering Spectral Clustering Active Learning	Three-way Clustering
Yu et al. (2020b)	Internal	Accuracy NMI	No	Three-way Clustering Active learning	Three-way Clustering
Zhu et al. (2020)	Internal	AUC Precision	t Test	Three-way Clustering	Multi-label

Table 4 Results for the weak supervision task: model optimization

Paper	Imputation	Feature selection	Hyperparameter optimization	Hyperparameters	Reported improvement
Campagner et al. (2020b)	No	No	No	Num. of estimators	1% to 25%
Campagner and Ciucci (2018)	N/A	N/A	N/A	N/A	N/A
(Campagner et al. 2021b)	No	No	No	Num. of estimators 2 thresholds	0% to 6%
Chen et al. (2015)	No	No	No	None	Not clearly reported
Dai et al. (2019)	No	PCA	Not reported	Num. of neighbors Num. of components	No comparison
Shengdan et al. (2022)	No	No	No	Num. of neighbors	−0.5% to 6.4%
Chenchen et al. (2017)	No	No	No	2 coefficients	Execution Time: 25%
Xiong and Yu (2019)	No	No	No	1 coefficient 1 coefficient per view Num. of clusters	Accuracy: 1% to 6% NMI: −1% to 6% ARI: −2% to 2%
Yu et al. (2017b)	No	No	Not reported	Num. of clusters 2 thresholds Num. of neighbors	Accuracy: 1% to 7% NMI: 2% to 4%
Yu et al. (2020b)	Not reported	No	Parameter study	2 coefficients Num. of clusters Num. of queries	Accuracy: −5% to 5% NMI: −5% to 6%
Zhu et al. (2020)	No	No	Not reported	7 coefficients 2 coefficients per view Num. of clusters	AUC: 0 to 8% Precision: 0% to 11%

Table 5 Results for the missing data task: general Information

Paper	Authors' affiliations	Task	Number of datasets	Datasets' domains	Datasets' sources	Dimensionality
Afridi et al. (2018)	Pakistan Canada Saudi Arabia	Missing data Clustering	4	Medicine Images (Digits) General	UCI	Instances: 150 to 10992 Features: 4 to 16 Classes: 2 to 10
Ali et al. (2021)	Pakistan Canada	Unsupervised Clustering Outlier detection	7	Scientific Medicine	UCI CHAMELEON	Instances: 768 to 8000 Attributes: 2 to 21 Classes: 2 to 31
Huang et al. (2013)	China	Missing data Rule induction	Theoretical Paper	N/A	N/A	N/A
Li et al. (2020)	China	Missing data Classification	Theoretical Paper	N/A	N/A	N/A
Luo et al. (2020a)	China Canada Japan	Missing data Rule induction	Theoretical Paper	N/A	N/A	N/A
Luo et al. (2020b)	China Canada	Missing data Rule induction	Theoretical Paper	N/A	N/A	N/A
Luo (2021)	China	Hybrid Data Classification	Theoretical Paper	N/A	N/A	N/A
Nowicki et al. (2020)	Poland Japan	Missing data Interval data Classification	3	Medicine General	UCI	Instances: 214 to 569 Features: 8 to 9 Classes: 2
Sakai et al. (2020)	Japan Malaysia	Missing data Rule induction	10 Theoretical Paper	Not reported	UCI Private[12]	Not reported
Wang et al. (2020)	China	Missing Data Imputation Classification	12	General	UCI[13]	Instances: 100 to 1502 Features: 2 to 4 Classes: 2 to 40
Wang and Chen (2020)	China	Missing data Imputation Clustering	6	Images (Digits) Text General	UCI	Instances: 150 to 5473 Features:4 to 30 Classes:2 to 10

Table 5 (continued)

Paper	Authors' affiliations	Task	Number of datasets	Datasets' domains	Datasets' sources	Dimensionality
Yang et al. (2020)	China Spain Japan	Interval-valued data Classification	6	Medicine Synthetic General	UCI	Instances: 150 to 1728 Features: 4 to 9 Classes: 2 to 4
Yang and Hou (2018)	China	Missing data Clustering	5	Synthetic Images (Satellite) General	UCI	Instances: 150 to 2000 Features: 2 to 36 Classes: 2 to 6
Yu (2017)	China	Missing data Clustering	3	General Text Images (Digits)	UCI	Instances: 150 to 10992 Features: 4 to 16 Classes: 3 to 10
Yu et al. (2014a)	China	Missing data Clustering	5	Synthetic General	UCI	Instances: 150 to 4839 Features: 2 to 7
Zhang et al. (2021)	China	Unsupervised Missing data Clustering	5	General Numbers	UCI	Instances: 150 to 5473 Features: 4 to 21 Clusters: 3 to 5

Table 6 Results for the missing data task: experimental setting

Paper	Validation Type	Evaluation Metrics	Significance Testing	Proposed TW methodology	Output type
Afridi et al. (2018)	Internal	Accuracy Coverage	No	Game-theoretic Rough Sets Three-way Clustering	Three-way Clustering
Ali et al. (2021)	Internal	Accuracy Precision Recall F1-score Davies–Bouldin index Average Silhouette index Outliers detection	t-Test	Three-way Clustering Set-pair analysis	Three-way Clustering
Huang et al. (2013)	N/A	N/A	N/A	Interval Algebra Rough Sets	Set of Rules
Li et al. (2020)	N/A	N/A	N/A	Tolerance-based Decision Theoretic Rough Sets	Single label with abstention
Luo et al. (2020a)	N/A	N/A	N/A	Similarity-based Rough Sets Logics	Set of Rules
Luo et al. (2020b)	N/A	N/A	N/A	Rough Sets Fuzzy Logic	Set of Rules
Luo (2021)	N/A	N/A	N/A	Decision Theoretic Rough Sets	Set of Rules
Nowicki et al. (2020)	Cross-validation	Accuracy Coverage	No	Rough Sets SVM Interval Analysis	Single class with abstention
Sakai et al. (2020)	N/A	N/A	N/A	Rough Sets Apriori algorithm Possible-world semantics	Set of Rules
Wang et al. (2020)	Internal	Accuracy	No[14]	Active Learning Ensemble Imputation	Imputed values

Table 6 (continued)

Paper	Validation Type	Evaluation Metrics	Significance Testing	Proposed TW methodology	Output type
Wang and Chen (2020)	Internal	Accuracy Fowlkes–Mallows index	No	Three-way Clustering Ensemble Clustering Cluster-based Imputation	Three-way Clustering
Yang et al. (2020)	Internal	Error rate	No	Dominance-based Variable Precision Rough Sets	Single label with abstention
Yang and Hou (2018)	Internal	Accuracy	No	Three-way clustering Density-based clustering	Three-way Clustering
Yu (2017)	Internal	Accuracy F1 score	Confidence intervals	Three-way Clustering	Three-way Clustering
Yu et al. (2014a)	Internal	Accuracy	No	Cluster-based Imputation Three-way Clustering	Three-way Clustering
Zhang et al. (2021)	Internal	Accuracy	No	Three-way Clustering Set-pair analysis	Three-way Clustering

Table 7 Results for the missing data task: model optimization

Paper	Imputation	Feature selection	Hyperparameter optimization	Hyperparameters	Reported improvement
Afridi et al. (2018)	No	No	No	Num. of clusters 4 thresholds Num. of iterations	Accuracy: 4% to 17% Coverage: Not reported
Ali et al. (2021)	No	No	No	Set of clusters	Accuracy: -5% to 2% Precision: -17% to 5% Recall: -29% to 1% F1-score: -7% to 2% Davies–Bouldin index: 0.73% to 11.35% Average Silhouette index: 0.53% to 9.51% Outlier detection: 2.5% to 21.3%
Huang et al. (2013)	N/A	N/A	N/A	N/A	N/A
Li et al. (2020)	N/A	N/A	N/A	N/A	N/A
Luo et al. (2020a)	N/A	N/A	N/A	N/A	N/A
Luo et al. (2020b)	N/A	N/A	N/A	N/A	N/A
Luo (2021)	N/A	N/A	N/A	N/A	N/A
Nowicki et al. (2020)	No	No	Not reported	Kernel function	No comparison
Sakai et al. (2020)	N/A	N/A	N/A	N/A	N/A
Wang et al. (2020)	Proposed method	No	No	Num. of Neighbors	-4.75% to 20%
Wang and Chen (2020)	Proposed Method	No	Not reported	Num. of clusters Ensemble size	Accuracy: -3% to 22% FMI: -1% to 31%
Yang et al. (2020)	No	No	Yes	2 decision thresholds 1 precision parameter	0.2% to 1.3%
Yang and Hou (2018)	No	No	Not reported	Num. of clusters Distance threshold	0% to 17%

Table 7 (continued)

Paper	Imputation	Feature selection	Hyperparameter optimization	Hyperparameters	Reported improvement
Yu (2017)	No	No	No	Num. of clusters 2 thresholds Radius threshold	Accuracy: 1% to 11% F1: 0% to 10%
Yu et al. (2014a)	Proposed Method	No	No	Num. of neighbors Num. of important features Density threshold Radius threshold 2 thresholds	No comparison
Zhang et al. (2021)	Not reported	No	Not reported	Num. of clusters 2 thresholds	Accuracy: -0.70% to 20.80%

Table 8 Results for the classification task: general Information

Paper	Authors' affiliations	Task	Number of datasets	Datasets' domains	Datasets' sources	Dimensionality
Campagner et al. (2021)	Italy	Supervised Conformal prediction	12	Medicine General	UCI	Instances: 178 to 581012 Features: 8 to 130107 Classes: 2 to 20
Campagner et al. (2019a)	Italy	Classification	1	Medicine	IRCCS galeazzi (Private)	Instances: 462 Features: 9 Classes: 2
Campagner et al. (2019b)	Italy	Classification	7	Medicine Biology Images (digits) Images (faces) General	IRCCS galeazzi (Private) UCI	Instances: 150 to 1797 Features: 4 to 4096 Classes: 2 to 40
Campagner et al. (2020b)	Italy	Partial labels Classification	7	Medicine Images (digits) Images (faces) General	IRCCS galeazzi (Private) UCI	Instances: 150 to 1797 Features: 4 to 4096 Classes: 2 to 40
Chen et al. (2020)	China	Keyword extraction	2	Text (articles)	GitHub[15]	2000 documents
Chen et al. (2016)	China	Feature selection Classification	6	Medicine Biology General	UCI	Instances: 100 to 768 Features: 7 to 19 Classes: 2 to 8
Deng and Jia (2016)	China	Classification	10	Biology General	UCI	Instances: 625 to 67557 Features: 4 to 60 Classes: 3 to 26
Jia et al. (2019)	China	Classification	10	General	UCI	Instances: 625 to 67557 Features: 4 to 60 Classes: 3 to 26
Li et al. (2017a)	China	Concept mining	5 Theoretical paper	Biology General	UCI	Instances: 20000 to 1025010 Features: 4 to 23

Table 8 (continued)

Paper	Authors' affiliations	Task	Number of datasets	Datasets' domains	Datasets' sources	Dimensionality
Li et al. (2013)	China	Classification	18	Medicine Finance Politics General	UCI	Instances: 106 to 2126 Classes: 2 to 15
Li et al. (2019)	China USA	Feature Selection Classification	15	Images (digits) Medicine Biology Text General	UCI	Instances: 80 to 8124 Features: 6 to 857
Li et al. (2017b)	Australia Canada Hong Kong	Classification	2	Text (general)	UCI	Instances: 21578 to 111740 Classes:10 to 50
Liang and Yi (2021)	China	Classification	2	Texts (policy)	Private (China govt.)	Classes: 5
Liu et al. (2015)	China	Classification	Theoretical Paper	N/A	N/A	N/A
Nauman et al. (2016)	Pakistan Canada	Classification	1	Software	UNM[16]	Not reported
Qian et al. (2021)	China	Multi-label Classification Feature selection	15	Biology Medicine Text (news) General	Mulan[17] Weka[18]	Instances: 194 to 7766 Features: 16 to 1000 Classes: 6 to 174
Singh and Rabadiya (2018)	India	Classification	Theoretical Paper	N/A	N/A	N/A
Singh et al. (2021)	Canada	Classification	1	Medicine	University of California Guangzhou Medical Center	Instances: 5878 Classes: 2
Subhashini et al. (2020)	Sri Lanka Australia	Classification Opinion Mining	2	Text (reviews)	MR[19]	Not reported
Yue et al. (2021)	China	Classification	2	Medicine	Breast IDC[20]	Instances: 2838 to 155314 Classes: 2

Table 8 (continued)

Paper	Authors' affiliations	Task	Number of datasets	Datasets' domains	Datasets' sources	Dimensionality
Zhang et al. (202)	China Canada Poland	Multi-label Classification	6	Biology Text (general)	Mulan Meka	Instances: 662 to 7395 Features: 1001 to 1836 Classes: 22 to 159
Zhang et al. (2019a)	China	Classification	5	Text (reviews) Text (general)	IMDb MR CR[21] MPQA[22]	Instances: 3780 to 50000 Classes: 2
Zhang et al. (2018)	China	Multi-label Classification	6	Biology Text (general)	Mulan Meka	Instances: 662 to 7395 Features: 1001 to 1836 Classes: 22 to 159
Zhang et al. (2019)	China	Classification	4	General	MR CR SUBJ[23] MPQA[24]	Instances: 3780 to 10662
Zhou and Yao (2011)	Canada	Decision support Classification	Theoretical paper	N/A	N/A	N/A
Zhou et al. (2014)	USA Canada	Classification	3	Text	UCI Ling Spam[25] PU1 Deshpande et al. (2007)	Instance: 1099 to 4601 Classes: 2 Features: 58 to 300
Zhu et al. (2014)	China	Classification Sentence recognition	Not reported	Not reported	Not reported	Not reported

Table 9 Results for the classification task: experimental setting

Paper	Validation type	Evaluation metrics	Significance testing	Proposed TW methodology	Output type
Campagner et al. (2021)	Cross-validation	Accuracy Coverage	Confidence interval; Friedman Test; Nemeny Test	Three-way Cautious learning	Set of classifiers
Campagner et al. (2019a)	Nested cross-validation	Accuracy Balanced accuracy	No	Decision tree	Single label with abstention
Campagner et al. (2019b)	Nested cross-validation	Accuracy	Friedman test Confidence intervals	Decision tree Ensemble learning Belief functions Convex optimization	Single label with abstention
Campagner et al. (2020b)	Cross-validation	Accuracy	Friedman Test	Decision tree Ensemble learning Belief function Optimization-based	Single label with abstention
Chen et al. (2020)	Hold-out	F1 score	No	Graph theory	List of keyphrases
Chen et al. (2016)	Cross-validation	Accuracy	No	Neighborhood systems Rough sets	Selected features
Deng and Jia (2016)	Cross-validation	Accuracy Abstention rate Cost	No	Decision-theoretic rough sets	Single label with abstention
Jia et al. (2019)	Repeated cross-validation	Accuracy Cost	T test	Decision-theoretic rough sets Multi-stage classification	Single label with abstention
Li et al. (2017a)	N/A	N/A	N/A	Formal concept analysis rough sets Granular computing	Set of Orthopairs
Li et al. (2013)	Cross-validation	Accuracy	No	Ensemble learning Multi-stage classification	Single label
Li et al. (2019)	Repeated cross-validation	Accuracy	Confidence intervals	Decision-theoretic rough sets Ensemble learning Information theory Multi-objective Optimization	Selected features

Table 9 (continued)

Paper	Validation type	Evaluation metrics	Significance testing	Proposed TW methodology	Output type
Li et al. (2017b)	Hold-out	F1 score Accuracy AUC	Wilcoxon signed-rank test	Multi-stage classification	Single label
Liang and Yi (2021)	Hold-out	Accuracy Macro-F1	No	Ensemble learning Multi-stage classification	Single label
Liu et al. 2015)(	N/A	N/A	N/A	Decision tree	Single label with abstention
Nauman et al. (2016)	Cross-validation	Accuracy Generality	No	Game-theoretic rough sets Information-theoretic rough sets	Single label with abstention
Qian et al. (2021)	Cross-validation	Hamming Loss Ranking Loss One Error Subset accuracy Average precision Average recall	Friedman test T test	Neighborhood-based Rough Sets	Multi-label
Singh and Rabadiya (2018)	N/A	N/A	N/A	Rough sets Four-way Decision space	Set of Rules
Singh et al. (2021)	Cross-validation	Accuracy Coverage Precision Recall True negative rate	No	Game-theoretic rough sets	Single label with abstention
Subhashini et al. (2020)	Internal	F1-score	t-Test (Two tailed)	Fuzzy formal concept analysis	Set of terms
Yue et al. (2021)	Hold-out	Accuracy F1 score Precision Recall Cost	No	Deep learning Evidence theory	Single label with abstention

Table 9 (continued)

Paper	Validation type	Evaluation metrics	Significance testing	Proposed TW methodology	Output type
Zhang et al. (2020)	Cross-validation	Execution Time[26] Hamming loss Precision Recall Micro-F1	No	Decision-theoretic rough sets Ensemble learning Granular computing	Multi-label
Zhang et al. (2019a)	Repeated hold-out	Accuracy	T test Confidence intervals	Ensemble learning Multi-stage classification	Single label
Zhang et al. (2018)	Repeated hold-out	Precision Recall Accuracy F1 score Micro-F1 Hamming loss	Friedman test Confidence interval	Decision-theoretic rough sets Ensemble learning	Multi-label
Zhang et al. (2019)	Cross-validation	Accuracy	No[27]	Deep learning Sequential	Single label
Zhou and Yao (2011)	N/A	N/A	N/A	Ternary classification Rough sets	Decision Tree
Zhou et al. (2014)	Cross-validation	Precision Recall Weighted accuracy Weighted error Cost	No	Decision-theoretic rough sets Robinson algorithm Robinson (2003)	Single label with abstention
Zhu et al. (2014)	Internal	Precision Recall F1-score	No	Decision-theoretic rough sets	Multi-label

Table 10 Results for the classification task: model optimization

Paper	Imputation	Feature selection	Hyperparameter optimization	Hyperparameters	Reported improvement
Campagner et al. (2021)	No	No	No	Num. of neighbors Num. of decision trees	Accuracy Random Forest: 0% to 3% Accuracy kNN: -2% to 8% Coverage Random Forest: -20% to 0% Coverage kNN: -40% to 4%
Campagner et al. (2019a)	No	No	Nested Cross-validation	2 thresholds Tree depth	Accuracy: 2% Balanced Accuracy: 10%
Campagner et al. (2019b)	No	No	Nested Cross-validation	Cost matrix	0% to 22%
Campagner et al. (2020b)	No	No	No	Num. of estimators Cost matrix 1 thresholds	1% to 24%
Chen et al. (2020)	No	No	Parameter study	1 threshold	1.4% to 6.6%
Chen et al. (2016)	No	Proposed method	No	3 thresholds	-0.9% to 4%
Deng and Jia (2016)	No	No	No	Cost matrix 2 coefficients	Accuracy: 0.7% to 3.7% Abstention rate: -22.7% to -5.9% Cost: 0.7 to 530
Jia et al. 2019)(	No	No	Parameter study	Cost matrix 2 coefficients	Accuracy: -2.8% to 16% Cost: -11% to 7%
Li et al. (2017a)	N/A	N/A	N/A	2 thresholds	N/A
Li et al. (2013)	No	No	No	Cost matrix Num. of estimators	0.1% to 6.7%
Li et al. (2019)	Yes Method not reported	Proposed method	No	Cost matrix Num. of generations Population size	Accuracy: -3% to 4%
Li et al. (2017b)	No	No	Parameter study	4 thresholds	F1: 3.2% to 8.7% Accuracy: -0.7% to 0.6% AUC: 3.1% to 6.3%

Table 10 (continued)

Paper	Imputation	Feature selection	Hyperparameter optimization	Hyperparameters	Reported improvement
Liang and Yi (2021)	No	No	Parameter study	1 threshold Num. of estimators Learning rate Drop-out rate	Accuracy: 3% to 8% Macro-F1: 3% to 7%
Liu et al. 2015)(	N/A	N/A	N/A	N/A	N/A
Nauman et al. (2016)	No	No	No	2 thresholds	Accuracy: -1.8% to 10% Generality: -2% to 5.4%
Qian et al. (2021)	No	Proposed method	No	Neighborhood threshold 2 thresholds	Hamming Loss: -0.3% to 3% Ranking Loss: -1.3% to 1.5% One Error: -2% to 2% Subset Accuracy: -3% to 6% Average Precision: -2% to 2% Average Recall: -3% to 5%
Singh and Rabadiya (2018)	N/A	N/A	N/A	N/A	N/A
Singh et al. (2021)	No	No	Proposed method	2 thresholds	Accuracy: -2.4% Coverage: -36% Recall: 0.7% True Negative Rate: -4.6%
Subhashini et al. (2020)	No	bm25 UNI ICF	Standard derivation	1 coefficient 1 threshold	Precision: 2.11% to 6.02% Recall: 2.82% to 7.56% F1-score: 2.47% to 7.58%
Yue et al. (2021)	No	No	Not reported	1 coefficient 1 threshold	Accuracy: 1.1% F1 score: 1.6% Precision: -1.9% Recall: +3% Cost: -4.2
Zhang et al. (2020)	No	Proposed method	Parameter study	2 thresholds Num. of estimators Set of instances	Execution Time: 0.01s to 500s[28]

Table 10 (continued)

Paper	Imputation	Feature selection	Hyperparameter optimization	Hyperparameters	Reported improvement
Zhang et al. (2019a)	No	No	No	Cost matrix [29]	1% to 4%
Zhang et al. (2018)	No	Proposed method	Parameter study	4 thresholds Num. of estimators	Label-based Precision: 9.5% to 23.7% Example-based Precision: -7.5% to 23.2% Label-based Recall: -21.8& to 21.4% Example-based Recall: -23.4% to 0.2% Label-based Accuracy: -7.2% to 24.6% Example-based Accuracy: -9.4% to 4.5% Label-based F1: -43.4% to 5.6% Example-based F1: -12% to 5.2% Micro-F1: -9% to 10% Hamming Loss: -13% to 0.9%
Zhang et al. (2019)	No	No	Not reported	Not reported	Accuracy: -0.7& to 0.8%
Zhou and Yao (2011)	N/A	N/A	N/A	N/A	N/A
Zhou et al. (2014)	No	Information gain	No	3 thresholds	Not reported
Zhu et al. (2014)	No	Information gain	No	2 thresholds	Precision: -2.57% to 9.80% Recall: -11.06% to 8.27% F1-score: -0.06% to 4.72%

Table 11 Results for the clustering task: general information

Paper	Authors' affiliations	Task	Number of datasets	Datasets' domains	Datasets' sources	Dimensionality
Afridi et al. (2020)	Pakistan Canada	Clustering Determination of thresholds	5	Medicine General	UCI	Not reported
Jia et al. (2021)	China	Unsupervised Clustering	15	Medicine General	UCI	Instances: 32 to 625 Attributes: 4 to 57 Classes: 2 to 22
Jiang et al. (2019)	China Canada	Clustering Label matching	Theoretical Paper	N/A	N/A	N/A
Shi et al. (2018)	China	Clustering	5	General	UCI	Instances: 150 to 1372 Features: 4 to 100 Classes: 2 to 3
Sun and Yu (2018)	China	Number of clusters determination	6	General Images(Letters)	UCI	Instances: 150 to 528 Clusters: 3 to 11
Wang et al. (2019a)	China	Clustering	8	General	UCI	Instances: 310 to 2008 Attributes: 4 to 100 Classes: 2 to 4
Wang et al. (2017)	China Canada	Clustering Label matching	5	Medicine Scientific	UCI	Instances: 310 to 1372 Attributes: 4 to 100 Classes: 2
Wang et al. (2019)	China	Clustering	9	General Images(Digits)	UCI USPS	Instances: 178 to 20560 Attributes: 4 to 256 Classes: 2 to 6
Wang and Yang (2021)	China	Clustering	6	General Medicine Scientific	UCI	Instances: 106 to 19020 Attributes: 4 to 16 Classes: 2 to 6
Wang and Yao (2018)	China Canada	Clustering	8	Synthetic General	Private UCI USPS	Instances: 150 to 2420 Attributes: 4 to 256 Classes: 2 to 6
Yu (2018)	China	Clustering Cluster analysis	Theoretical paper	N/A	N/A	N/A

Table 11 (continued)

Paper	Authors' affiliations	Task	Number of datasets	Datasets' domains	Datasets' sources	Dimensionality
Yu et al. (2020)	China	Clustering	7	Medicine General Images(Letters)	UCI	Instances: 150 to 5000 Features: 4 to 44 Clusters: 2 to 7
Yu et al. (2017a)	China USA	Clustering for Mixed-type data	5	General	UCI	Instances: 150 to 48842 Attributes: 0 to 16 Clusters: 2 to 3
Yu et al. (2019a)	China Canada	Clustering Decision Support	10	Synthetic General	Private UCI	Instances: 210 to 5000 Features: 2 to 9 Clusters: 2 to 15
Yu et al. (2019)	China Canada	Clustering	19	Medicine General Images(Letters) Images(Digits)	UCR UCI	Instances: 60 to 1048576 Features: 3 to 3572 Classes: 2 to 11
Yu and Wang (2018)	China	Clustering	10	Synthetic General	UCI Private	Instances: 1555 to 15000 Attributes: 2 to 36 Clusters: 2 to 7
Yu et al. (2014b)	China	Incremental clustering	1 Theoretical paper	Synthetic	Private	Instances: 1000
Yu and Zhou (2013)	China	Clustering	14	Synthetic Scientific General	UCI	Instances: 101 to 11092 Attributes: 4 to 36 Clusters: 2 to 10

Table 12 Results for the clustering task: experimental setting

Paper	Validation type	Evaluation metrics	Significance testing	Proposed TW methodology	Output type
Afridi et al. (2020)	Internal	Accuracy Generality Davies–Bouldin index Average Silhouette index Algorithm's accuracy F1-score Hamming loss Precision Recall	No	Three-way clustering variance based	Three-way clustering
Jia et al. (2021)	Internal	Accuracy NMI	No	Three-way clustering; Sample similarity-based	Three-way clustering
Jiang et al. (2019)	N/A	N/A	N/A	Three-way clustering	Three-way clustering
Shi et al. (2018)	Internal	Accuracy Davies–Bouldin index Average Silhouette Coefficient	No	Spectral clustering Three-way clustering	Three-way clustering
Sun and Yu (2018)	Internal	Number of predicted clusters	No	Three-way clustering Multi-validity index based	Number of clusters
Wang et al. (2019a)	Internal	Davies–Bouldin index Average Silhouette index Accuracy	No	Three-way clustering Cluster ensemble	Three-way clustering
Wang et al. (2017)	Internal	Micro-F1 Macro-F1	No	Hard clustering Ensemble clustering	Three-way clustering
Wang et al. (2019)	Internal	Davies–Bouldin index Average Silhouette index Accuracy	No	Three-way clustering K-means	Three-way clustering
Wang and Yang (2021)	Internal	Accuracy Davies–Bouldin index Average Silhouette Index	No	Three-way clustering Ensemble clustering Stability-based	Three-way clustering
Wang and Yao (2018)	Internal	Davies–Bouldin index Accuracy	No	Three-way clustering Mathematics morphology	Three-way clustering

Table 12 (continued)

Paper	Validation type	Evaluation metrics	Significance testing	Proposed TW methodology	Output type
Yu (2018)	N/A	N/A	N/A	Three-way clustering Uncertain soft c lustering	Three-way clustering
Yu et al. (2020)	Internal	Accuracy F1-score Rand Index NMI	No	Three-way clustering	Three-way clustering
Yu et al. (2017a)	Internal	ARI Accuracy	No	Three-way clustering Mixed-type data	Three-way clustering
Yu et al. (2019a)	Internal	Accuracy F1-score NMI	No	Three-way clustering Density-based	Three-way clustering
Yu et al. (2019)	Internal	ARI Accuracy NMI F1-Measure Time	No	Three-way clustering Cluster ensemble Large-scale data	Three-way clustering
Yu and Wang (2018)	Internal	Accuracy Computational time	No	Three-way clustering Cluster ensemble	Three-way clustering
Yu et al. (2014b)	N/A	N/A	N/A	Soft clustering Searching tree	Three-way clustering
Yu and Zhou (2013)	Cross-validation	Accuracy	No	Three-way clustering Ensemble clustering	Three-way clustering

Table 13 Results for the clustering task: model optimization

Paper	Imputation	Feature Selection	Hyperparameter Optimization	Hyperparameters	Reported Improvement
Afridi et al. (2020)	No	No	Proposed method	2 Thresholds	Single Label Accuracy: -0.31% to 0.38% Single Label Generality: -20.73% to -7.79% Single Label Davies–Bouldin index: -17.95% to 0.50% Single Label Silhouette: -9.06% to -0.03% Multi Label Accuracy: 0.02% to 4.22% Multi Label Generality: -20.41% to -11.89% Multi Label Davies–Bouldin index: -13.43% to 11.15% Multi Label Silhouette: 3.85% to 6.03% Multi Label F1-score: 4.01% to 4.23% Multi Label Hamming Loss: 2.64% to 7.09% Overlapping region Precision: 24.15% to 31.91% Overlapping region Recall: -42.37% to -38.06% Overlapping region Accuracy: -0.84% to 2.32% Algorithm's Accuracy: -11.61% to 4.18%
Jia et al. (2021)	No	No	No	Num. of clusters 2 thresholds	Accuracy: -19.80% to 28.64% NMI: -0.85% to 2.46%
Jiang et al. (2019)	N/A	N/A	N/A	N/A	N/A
Shi et al. (2018)	No	No	No	Num. of clusters 3 parameters	Accuracy: 0.12% to 1.95% Davies–Bouldin index: -5.64% to 11.58% Average Silhouette Coefficient: -2.52% to 6.22%
Sun and Yu (2018)	No	No	No	Num. of clusters 2 thresholds	No comparison
Wang et al. (2019a)	No	No	No	Num. of thresholds	No comparison
Wang et al. (2017)	No	No	No	Num. of clusters	Micro-F1: 0.31% to 10.01% Macro-F1: 0.31% to 10.25%
Wang et al. (2019)	No	No	No	1 Thresholds	Davies–Bouldin index: -18% to 58% Silhouette: -2% to 21% Accuracy: -12% to 10%

Table 13 (continued)

Paper	Imputation	Feature Selection	Hyperparameter Optimization	Hyperparameters	Reported Improvement
Wang and Yang (2021)	No	No	No	Num. of clusters 1 threshold 1 parameter	Accuracy: -0.70% to 6.50% Davies–Bouldin index: 2.08% to 59.39% Average Silhouette Index: 0.50% to 5.96%
Wang and Yao (2018)	No	No	No	2 Coefficients Num. of neighbors	Davies–Bouldin index: -2.49% to 16.93% Accuracy: 1.33 % to 4.91%
Yu (2018)	N/A	N/A	N/A	N/A	N/A
Yu et al. (2020)	No	No	No	Num. of neighbors 2 thresholds	Accuracy: 0.52% to 70.22% F1-score: -3.28% to 29.24% Rand Index: 0.97% to 51.85% NMI: -1.86% to 14.64%
Yu et al. (2017a)	No	No	No	2 Thresholds Num. of neighbors	ARI: -2% to 25% Accuracy: -1% to 0.5%
Yu et al. (2019a)	No	No	No	1 Threshold	Accuracy: LB(-39% to 31%); UB(-1% to 9%) F1-score: LB(-25% to 27%); UB(-1% to 4%) NMI: LB(-25% to 34%); UB(-2% to 19%)
Yu et al. (2019)	No	No	No	2 Thresholds	ARI: -22% to 76% Accuracy: -48% to 50% NMI: -7% to 25% F1-Measure: -56% to 33% Time: 40x faster on avg
Yu and Wang (2018)	No	No	No	2 Thresholds	Accuracy: 0% to 50% Computational time: 1.5 to 10x faster
Yu et al. (2014b)	No	No	No	1 Parameter 2 Thresholds	N/A
Yu and Zhou (2013)	No	No	No	1 Threshold	Accuracy: -2% to 13%

Author contributions A.C. and F.M wrote the main manuscript text, A.C. performed the statistical analysis and prepared figures. A.C., D.C. and F.C. contributed to conceptual framework and definitions for the research. D.C. and F.C. supervised the research. All authors reviewed and approved the manuscript.

Declarations

Competing interests The authors declare no competing interests

References

Afridi MK, Azam N, Yao J, Alanazi E (2018) A three-way clustering approach for handling missing data using gtrs. Int J Approx Reason 98:11–24

Afridi MK, Azam N, Yao J (2020) Variance based three-way clustering approaches for handling overlapping clustering. Int J Approx Reason 118:47–63. https://doi.org/10.1016/j.ijar.2019.11.011

Afyouni I, Al Aghbari Z, Razack RA (2022) Multi-feature, multi-modal, and multi-source social event detection: a comprehensive survey. Inform Fusion 79:279–308

Ali B, Azam N, Shah A, Yao J (2021) A spatial filtering inspired three-way clustering approach with application to outlier detection. Int J Approx Reason 130:1–21. https://doi.org/10.1016/j.ijar.2020.12.003

Assent I (2012) Clustering high dimensional data. Wiley Interdiscip Rev 2(4):340–350

Bello R, Falcon R (2017) Rough sets in machine learning: a review. Thriving Rough Sets 8:87–118

Benavoli A, Corani G, Mangili F (2016) Should we really use post-hoc tests based on mean-ranks? J Mach Learn Res 17(1):152–161

Berrar D (2017) Confidence curves: an alternative to null hypothesis significance testing for the comparison of classifiers. Mach Learn 106:911–949

Boyd KL (2021) Datasheets for datasets help ml engineers notice and understand ethical issues in training data. In: Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2):1–27

Bussola N, Marcolini A, Maggio V, Jurman G, Furlanello C (2019) Not again! data leakage in digital pathology. arXiv preprint arXiv:1909.06539

Cabitza F, Campagner A (2021) The need to separate the wheat from the chaff in medical informatics: introducing a comprehensive checklist for the (self)-assessment of medical ai studies. Int J Med Inform 153:104510

Cabitza F, Campagner A, Soares F, de Guadiana-Romualdo LG, Challa F, Sulejmani A, Seghezzi M, Carobene A (2021) The importance of being external. methodological insights for the external validation of machine learning models in medicine. Comput Methods Programs Biomed 208:106288

Campagner A., Cabitza F, Ciucci D (2019a) Exploring medical data classification with three-way decision trees. In HEALTHINF 2019-12th International Conference on Health Informatics. In: Proceedings; Part of 12th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2019, pp 147–158. URL www.scopus.com

Campagner A, Cabitza F, Ciucci D (2019b) Three–way classification: Ambiguity and abstention in machine learning. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11499 LNAI:280–294. URL www.scopus.com

Campagner A, Ciucci D (2018) Three-way and semi-supervised decision tree learning based on orthopartitions. In: Medina J, Ojeda-Aciego M, Verdegay JL, Pelta DA, Cabrera IP, Bouchon-Meunier B, Yager RR (eds) Theory and foundations. Springer International Publishing, Cham, pp 748–759

Campagner A, Ciucci D (2019) Orthopartitions and soft clustering: soft mutual information measures for clustering validation. Knowl-Based Syst 180:51–61

Campagner A, Cabitza F, Ciucci D (2020a) Three-way decision for handling uncertainty in machine learning: a narrative review. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12179 LNAI:137–152
Campagner A, F Cabitza, D Ciucci (2020b) The three-way-in and three-way-out framework to treat and exploit ambiguity in data. Int J Approx Reason, 119:292 – 312. ISSN 0888-613X. https://doi.org/10.1016/j.ijar.2020.01.010
Campagner A, Cabitza F, Berjano PL, Ciucci D (2021) Three-way decision and conformal prediction: isomorphisms, differences and theoretical properties of cautious learning approaches. Inform Sci 579:347–367. https://doi.org/10.1016/j.ins.2021.08.009
Campagne A, Ciucci D, Svensson CM, Figge MT, Cabitza F (2021) Ground truthing from multi-rater labeling with three-way decision and possibility theory. Inform Sci 545:771–790. https://doi.org/10.1016/j.ins.2020.09.049
Campagner A, Ciucci D, Denœux T (2022) Belief functions and rough sets: survey and new insights. Int J Approx Reason 143:192–215
Campagner A, Ciucci D, Denœux T (2022b) A distributional approach for soft clustering comparison and evaluation. In: International Conference on Belief Functions, pp 3–12. Springer
Campagner A, Ciucci D, Denœux T (2023) A distributional framework for evaluation, comparison and uncertainty quantification in soft clustering. Int J Approx Reason 162:109008
Campagner A, Ciucci D, Denœux T (2023) A general framework for evaluating and comparing soft clusterings. Inform Sci 623:70–93
Chao G, Sun S, Bi J (2017) A survey on multi-view clustering. arXiv preprint arXiv:1712.06246
Chen J, Zhao S, Yanping Z (2015) A multi-view decision model based on cca. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, 9436 LNAI:266–274. https://doi.org/10.1007/978-3-319-25754-9_24
Chen T, Miao D, Zhang Y (2020) A graph-based keyphrase extraction model with three-way decision. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12179 LNAI:111–121. URL www.scopus.com
Chen Y, Zeng Z, Zhu Q, Tang C (2016) Three-way decision reduction in neighborhood systems. Appl Soft Comput J 38:942–954
Chen YT, Witten DM (2022) Selective inference for k-means clustering. arXiv preprint arXiv:2203.15267
Crossnohere NL, Elsaid M, Paskett J, Bose-Brill S, Bridges JFP (2022) Guidelines for artificial intelligence in medicine: literature review and content analysis of frameworks. J Med Internet Res 24(8):e36823
Dai D, Zhou X, Li H, Liu L (2019) Co-training based sequential three-way decisions for cost-sensitive classification. In: 2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC), pp 157–162.https://doi.org/10.1109/ICNSC.2019.8743205
Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30
Deng G, Jia X (2016) A decision-theoretic rough set approach to multi-class cost-sensitive classification. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9920 LNAI:250–260. URL www.scopus.com
Denoeux T, Li S, Sribonchitta S (2017) Evaluating and comparing soft partitions: An approach based on dempster-shafer theory. IEEE Trans Fuzzy Syst 26(3):1231–1244
Deshpande VP, RF Erbacher, C Harri (2007) An evaluation of naïve bayesian anti-spam filtering techniques. In: 2007 IEEE SMC Information Assurance and Security Workshop, pp 333–340. IEEE
Destercke S. (2022) Uncertain data in learning: challenges and opportunities. In: U Johansson, H Boström, KA Nguyen, Z Luo, and L Carlsson, (eds.), In: Proceedings of the eleventh symposium on conformal and probabilistic prediction with applications, volume 179 of Proceedings of Machine Learning Research, pp 322–332. PMLR
Dodge J, Gururangan S, Card D, Schwartz R (2019) Show your work: improved reporting of experimental results. arXiv preprint arXiv:1909.03004
Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O (2021) A survey on missing data in machine learning. J Big Data 8(1):1–37
Gao LL, Bien J, Witten D (2022) Selective inference for hierarchical clustering. J Am Stat Ass 8:1–27
García-Pérez MA (2023) Use and misuse of corrections for multiple testing. Methods Psychol 82023:100120
Golfarelli M, Maio D, Malton D (1997) On the error-reject trade-off in biometric verification systems. IEEE Trans Pattern Anal Mach Intell 19(7):786–796
Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG (2016) Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 31:337–350
Hansen LK, Liisberg C, Salamon P (1997) The error-reject tradeoff. Open Syst Inform Dyn 4(2):159–184
Hannes H, Nelson C (2020) Building machine learning pipelines. O'Reilly Media

Hendrickx K, Perini L, Van der Plas D, Meert W, Davis J (2021) Machine learning with a reject option: a survey. arXiv preprint arXiv:2107.11277

Huang C, Li J, Wu WZ (2017) An information fusion viewpoint: three-way concept learning based on cognitive operators. Int J Approx Reason 83:218–242. https://doi.org/10.1016/j.ijar.2017.01.009

Huang S, Wang Q, Cheng J, Wu Z (2013) A semantic interpretation of rules in interval sets. In: Proceedings-International Conference on Natural Computation, pp 1000–1004. URL www.scopus.com

Hüllermeier E, Waegeman W (2021) Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach Learn 110(3):457–506

Hullermeier E, Rifqi M, Henzgen S, Senge R (2011) Comparing fuzzy partitions: a generalization of the rand index and related measures. IEEE Trans Fuzzy Syst 20(3):546–556

Japkowicz N (2013) Assessment metrics for imbalanced learning. Imbalanced learning: Foundations, algorithms, and applications, pp 187–206

Jia X, Li W, Shang L (2019) A multiphase cost-sensitive learning method based on the multiclass three-way decision-theoretic rough set model. Inform Sci 485:248–262

Jia X, Rao Y, Li W, Yang S, Yu H (2021) An automatic three-way clustering method based on sample similarity. Int J Mach Learn Cybernet 12(5):1545–1556. https://doi.org/10.1007/s13042-020-01255-8

Jiang C, Duan Y, Yao J (2019) Resource-utilization-aware task scheduling in cloud platform using three-way clustering. J Intell Fuzzy Syst 37(4):5297–5305. https://doi.org/10.3233/JIFS-190459

Kapoor S, Arvind N (2022) Leakage and the reproducibility crisis in ml-based science. arXiv preprint arXiv:2207.07048

Kompa B, Snoek J, Beam AL (2021) Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digital Med 4(1):1–6

Lei Y, Bezdek JC, Roman S, Vinh NX, Chan J, Bailey J (2017) Ground truth bias in external cluster validity indices. Pattern Recognit 65:58–70

Lenz OU , D Peralta, C Cornelis(2022) No imputation without representation. arXiv preprint arXiv:2206.14254

Li J, Huang C, Qi J, Qian Y, Liu W (2017a) Three-way cognitive concept learning via multi-granularity. Inform Sci 378:244–263. URL www.scopus.com

Li W, Huang Z, Jia X (2013) Two-phase classification based on three-way decisions. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8171 LNAI:338–345. URL www.scopus.com

Li W, Jia X, Wang L, Zhou B (2019) Multi-objective attribute reduction in three-way decision-theoretic rough set model. Int J Approx Reason 105:327–341

Li Y, Zhang L, Xu Y, Yao Y, Lau RYK, Wu Y (2017b) Enhancing binary classification by modeling uncertain boundary in three-way decisions. IEEE Trans Knowl Data Eng 29(7):1438–1451. URL www.scopus.com

Li Z, Xie N, Huang D, Zhang G (2020) A three-way decision method in a hybrid decision information system and its application in medical diagnosis. Artif Intell Rev 53(7):4707–4736

Liang D, Yi B (2021) Two-stage three-way enhanced technique for ensemble learning in inclusive policy text classification. Inform Sci 547:271–288

Lienen J, Hüllermeier E (2021) Credal self-supervised learning. Adv Neural Inform Process Syst 34:89

Lingras P, West C (2004) Interval set clustering of web users with rough k-means. J Intell Inform Syst 23(1):5–16

Lipton ZC, Steinhardt J (2018) Troubling trends in machine learning scholarship. arXiv preprint arXiv:1807.03341

Little RJA, Rubin DB (2019) Statistical analysis with missing data, volume 793. Wiley

Liu Y, Xu J, Sun L, Du L (2015) Decisions tree learning method based on three-way decisions. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9437 LNAI:389–400. URL www.scopus.com

Luo J, Fujita H, Yao Y, Qin K (2020a) On modeling similarity and three-way decision under incomplete information in rough set theory. Knowl-Based Syst, 191. URL www.scopus.com

Luo J, Hu M, Qin K (2020b) Three-way decision with incomplete information based on similarity and satisfiability. Int J Approx Reason 120:151–183. URL www.scopus.com

Luo S (2021) A three-way decision method based on hybrid data. J Intell Fuzzy Syst 40(5):8639–8650. https://doi.org/10.3233/JIFS-182764

Ma M (2016) Advances in three-way decisions and granular computing. Knowl.-Based Syst. 912016:1–3

Matthew H (2018) Has artificial intelligence become alchemy?

McDermott Matthew BA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M (2021) Reproducibility in machine learning for health research: still a ways to go. Sci Transl Med 13(586):eabb1655

Miao Y, Gao Y, Guo S, Liu W (2018) Incomplete data management: a survey. Front Comput Sci 12(1):4–25
Mongeon P, Paul-Hus A (2016) The journal coverage of web of science and scopus: a comparative analysis. Scientometrics 106:213–228
Mortier T, Wydmuch M, Dembczyński K, Hüllermeier E, Waegeman W (2021) Efficient set-valued prediction in multi-class classification. Data Min Knowl Discov 35(4):1435–1469
Nadeem MSA, Zucker JD, Hanczar B (2009) Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. In: Machine Learning in Systems Biology, pp 65–81. PMLR
Nauman M, Azam N, Yao J (2016) A three-way decision making approach to malware analysis using probabilistic rough sets. Inform Sci 374:193–209
Nowicki RK, Grzanek K, Hayashi Y (2020) Rough support vector machine for classification with interval and incomplete data. J Artif Intell Soft Comput Res 10(1):47–56
Ojeda FM, Jansen ML, Thiéry A, Blankenberg S, Weimar C, Schmid M, Ziegler A (2023) Calibrating machine learning approaches for probability estimation: a comprehensive comparison. Stat Med 42(29):5451–5478
Olatz A, Ibai G, Javier M, Pérez Jesús M, Iñigo P (2013) An extensive comparative study of cluster validity indices. Pattern Recognit 46(1):243–256
Pawlak Z (1982) Rough sets. Int J Comput Inform Sci 11:341–356
Pawlak Z (1991) Rough sets: theoretical aspects of reasoning about data. Springer, New York
Pawlak Z, Skowron A (2007) Rough sets: some extensions. Inform Sci 177(1):28–40
Poyiadzi R, Bacaicoa-Barber D, Cid-Sueiro J, Perello-Nieto M, Flach P, Santos-Rodriguez R (2022) The weak supervision landscape. In: 2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), pp 218–223. IEEE
Pugliese R, Regondi S, Marini R (2021) Machine learning-based approach: Global trends, research directions, and regulatory standpoints. Data Sci Manag 4:19–29
Qian W, Huang J, Wang Y, Xie Y (2021) Label distribution feature selection for multi-label classification with rough set. Int J Approx Reason 128:32–55
Rendón E, Abundez I, Arizmendi A, Quiroz EM (2011) Internal versus external cluster validation indexes. Int J Comput Commun 5(1):27–34
Robinson G (2003) A statistical approach to the spam problem. Linux J 2003(107):3
Sakai H, Nakata M, Watada J (2020) Nis-apriori-based rule generation with three-way decisions and its application system in sql. Inform Sci 507:755–771
Shalev-Shwartz S, Ben-David S (2014) Understanding machine learning: from theory to algorithms. Cambridge university press, Cambridge
Shao W, He L, Yu PS (2015) Clustering on multi-source incomplete data via tensor modeling and factorization. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp 485–497. Springer
Shengdan H, Miao D, Pedrycz W (2022) Multi granularity based label propagation with active learning for semi-supervised classification. Expert Syst Appl 192:116276
Shi H, Liu Q, Wang P (2018) Three-way spectral clustering. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11177 LNAI:389–398. https://doi.org/10.1007/978-3-030-01851-1_37
Singh P, Rabadiya K (2018) Uncertain information classification: a four-way decision making approach. pp 100–108. Institute of Electrical and Electronics Engineers Inc..https://doi.org/10.1109/ICAPR.2017.8593087
Singh S, Yao JT(2021) Pneumonia detection with game-theoretic rough sets. In: 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), pp 1029–1034. IEEE
Steyerberg EW, Harrell FE (2016) Prediction models need appropriate internal, internal–external, and external validation. J Clin Epidemiol 69:245–247
Subhashini LDCS, Li Y, Zhang J, Atukorale AS (2020) Integration of fuzzy and deep learning in three-way decisions. volume 2020-November, pp 71–78. IEEE Computer Society. https://doi.org/10.1109/ICDMW51313.2020.00019
Sun N, Yu H (2018) A method to determine the number of clusters based on multi-validity index. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11103 LNAI:427–439. https://doi.org/10.1007/978-3-319-99368-3_33
Thelwall M, Sud P (2022) Scopus 1900–2020: growth in articles, abstracts, countries, fields, and journals. Quant Sci Stud 3(1):37–50
Trivedi S, Pardos ZA, Heffernan NT (2015) The utility of clustering in prediction tasks. arXiv preprint arXiv:1509.06163

Ullmann T, Hennig C, Boulesteix AL (2022) Validation of cluster analysis results on validation data: a systematic framework. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, p e1444

Wang M, Li B, Min F, Liu J, Wang M (2020) Ensemble active imputation for incomplete data. In: 2020 IEEE International Conference on Networking, Sensing and Control, ICNSC 2020. URL www.scopus.com

Wang P, Chen X (2020) Three-way ensemble clustering for incomplete data. IEEE Access 8:91855–91864

Wang P, Yang X (2021) Three-way clustering method based on stability theory. IEEE Access 9:33944–33953. https://doi.org/10.1109/ACCESS.2021.3057405

Wang P, Yao Y (2018) Ce3: a three-way clustering method based on mathematical morphology. Knowl-Based Syst 155:54–65. https://doi.org/10.1016/j.knosys.2018.04.029

Wang P, Liu Q, Yang X, Xu F (2017) Ensemble re-clustering: Refinement of hard clustering by three-way strategy. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10559 LNCS:423–430. https://doi.org/10.1007/978-3-319-67777-4_37

Wang P, Liu Q, Xu G, Wang K (2019a) A three-way clustering method based on ensemble strategy and three-way decision. Information (Switzerland), 10(2).https://doi.org/10.3390/info10020059

Wang P, Shi H, Yang X, Mi J (2019) Three-way k-means: integrating k-means and three-way decision. Int J Mach Learn Cybernet 10(10):2767–2777. https://doi.org/10.1007/s13042-018-0901-y

Williams D, Liao X, Xue Y, Carin L, Krishnapuram B (2007) On classification with incomplete data. IEEE Trans Pattern Anal Mach Intell 29(3):427–436

Xiong J, Yu H (2019) A three-way clustering algorithm via decomposing similarity matrices for multi-view data with noise. In: Rough Sets, pp 179–193, Cham . Springer International Publishing

Yang D, Deng T, Fujita H (2020) Partial-overall dominance three-way decision models in interval-valued decision systems. Int J Approx Reason 126:308–325

Yang L, Hou K (2018) A method of incomplete data three-way clustering based on density peaks. In AIP Conference Proceedings, volume 1967. URL www.scopus.com

Yao Y (2010) Three-way decisions with probabilistic rough sets. Inform Sci 180(3):341–353

Yao Y (2012) An outline of a theory of three-way decisions. In: Rough Sets and Current Trends in Computing, pp 1–17. Springer, Berlin

Yao Y (2018) Three-way decision and granular computing. Int J Approx Reason 103:107–123

Yao Y (2022) Symbols-meaning-value (smv) space as a basis for a conceptual model of data science. Int J Approx Reason 144:113–128

Yao Y, Lingras P, Wang R, Miao D (2009) Interval set cluster analysis: a re-formulation. In International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing, volume 5908 LNAI, pp 398–405. Springer

Yu H (2017) A framework of three-way cluster analysis. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10314 LNAI:300–312. URL www.scopus.com

Yu H, Wang G (2018) An efficient gradual three-way decision cluster ensemble approach. Commun Comput Inform Sci 854:711–723. https://doi.org/10.1007/978-3-319-91476-3_58

Yu H, Zhou Q (2013) A cluster ensemble framework based on three-way decisions. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8171 LNAI:302–312.https://doi.org/10.1007/978-3-642-41299-8_29

Yu H, Su T, Zeng X (2014a) A three-way decisions clustering algorithm for incomplete data. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8818 LNAI:765–776. URL www.scopus.com

Yu H, Zhang C, Hu F (2014b) An incremental clustering approach based on three-way decisions. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8536 LNAI:152–159.https://doi.org/10.1007/978-3-319-08644-6_16

Yu H, Chang Z, Zhou B (2017a) A novel three-way clustering algorithm for mixed-type data. pp 119–126. Institute of Electrical and Electronics Engineers Inc..https://doi.org/10.1109/ICBK.2017.38

Yu H, Chen L, Yao J, Wang X (2019a) A three-way clustering method based on an improved dbscan algorithm. Physica A: Statistical Mechanics and its Applications, 535. https://doi.org/10.1016/j.physa.2019.122289

Yu H, Chen Y, Lingras P, Wang G (2019) A three-way cluster ensemble approach for large-scale data. Int J Approx Reason 115:32–49. https://doi.org/10.1016/j.ijar.2019.09.001

Yu H, Chang Z, Wang G, Chen X (2020) An efficient three-way clustering algorithm based on gravitational search. Int J Mach Learn Cybernet 11(5):1003–1016. https://doi.org/10.1007/s13042-019-00988-5

Yu H (2018) Three-way decisions and three-way clustering. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11103 LNAI:13–28. https://doi.org/10.1007/978-3-319-99368-3_2
Yu H, Wang X, Wang G (2017b) A semi-supervised three-way clustering framework for multi-view data. In Rough Sets, pp 313–325, Cham. Springer International Publishing
Yu H, Wang X, Wang G, Zeng X (2020b) An active three-way clustering method via low-rank matrices for multi-view data. Information Sciences, 507:823–839. ISSN 0020-0255.https://doi.org/10.1016/j.ins.2018.03.009
Yue X, Chen Y, Yuan B, Lv Y (2021) Three-way image classification with evidential deep convolutional neural networks. Cognit Comput, pp 1–13
Zhang C, Gao R, Qin H, Feng R (2021) Three-way clustering method for incomplete information system based on set-pair analysis. Granul Comput 6(2):389–398. https://doi.org/10.1007/s41066-019-00197-z
Zhang Y, Miao D, Zhang Z, Xu J, Luo S (2018) A three-way selective ensemble model for multi-label classification. Int J Approx Reason 103:394–413
Zhang Y, Miao D, Wang J, Zhang Z (2019a) A cost-sensitive three-way combination technique for ensemble learning in sentiment classification. Int J Approx Reason, 105:85–97. URL www.scopus.com
Zhang Y, Miao D, Pedrycz W, Zhao T, Xu J, Yu Y (2020) Granular structure-based incremental updating for multi-label classification. Knowl-Based Syst 189:105066
Zhang Y, Zhang Z, Miao D, Wang J (2019) Three-way enhanced convolutional neural networks for sentence-level sentiment classification. Inform Sci 477:55–64
Zhou Z-H (2018) A brief introduction to weakly supervised learning. Nat Sci Rev 5(1):44–53
Zhou B, Yao Y (2011) search of effective granulization with dtrs for ternary classification. Int J Cognit Inform Nat Intell 5(3):47–60
Zhou B, Yao Y, Luo J (2010) A three-way decision approach to email spam filtering. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6085 LNAI:28–39
Zhou B, Yao Y, Luo J (2014) Cost-sensitive three-way email spam filtering. J Intell Inform Syst 42(1):19–45
Zhu X, Ghahramani Z, Lafferty JD (2002) Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02–107, Carnegie Mellon University
Zhu Y, H Tian, J Ma, J Liu, T Liang (2014) An integrated method for micro-blog subjective sentence identification based on three-way decisions and Naive Bayes. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8818 LNAI:844–855. https://doi.org/10.1007/978-3-319-11740-9_77
Zhu C, Ma L, Wang P, Miao D(2020) Multi-view and multi-label method with three-way decision-based clustering. In: Pattern Recognition and Computer Vision, pp 69–80, Cham. Springer International Publishing