Information Fusion 89 (2023) 228–253

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

A unifying view of class overlap and imbalance: Key concepts, multi-view panorama, and open avenues for research

Miriam Seoane Santos [a,*], Pedro Henriques Abreu [a], Nathalie Japkowicz [b], Alberto Fernández [c], João Santos [d,e]

[a] University of Coimbra, CISUC, Department of Informatics Engineering, Coimbra 3030-290, Portugal
[b] Department of Computer Science, American University, Washington, DC, 20016, USA
[c] Department of Computer Science and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence, DaSCI, University of Granada, Spain
[d] Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto, Portugal
[e] IPO-Porto Research Centre (CI-IPOP), Porto, Portugal

ARTICLE INFO

MSC:
00-01
99-00

Keywords:
Class imbalance
Imbalanced data
Class overlap
Data complexity
Data intrinsic characteristics
Complexity measures

ABSTRACT

The combination of class imbalance and overlap is currently one of the most challenging issues in machine learning. While seminal work focused on establishing class overlap as a complicating factor for classification tasks in imbalanced domains, ongoing research mostly concerns the study of their synergy over real-word applications. However, given the lack of a well-formulated definition and measurement of class overlap in real-world domains, especially in the presence of class imbalance, the research community has not yet reached a consensus on the characterisation of both problems. This naturally complicates the evaluation of existing approaches to address these issues simultaneously and prevents future research from moving towards the devise of specialised solutions. In this work, we advocate for a unified view of the problem of class overlap in imbalanced domains. Acknowledging class overlap as the overarching problem – since it has proven to be more harmful for classification tasks than class imbalance – we start by discussing the key concepts associated to its definition, identification, and measurement in real-world domains, while advocating for a characterisation of the problem that attends to multiple sources of complexity. We then provide an overview of existing data complexity measures and establish the link to what specific types of class overlap problems these measures cover, proposing a novel taxonomy of class overlap complexity measures. Additionally, we characterise the relationship between measures, the insights they provide, and discuss to what extent they account for class imbalance. Finally, we systematise the current body of knowledge on the topic across several branches of Machine Learning (Data Analysis, Data Preprocessing, Algorithm Design, and Meta-learning), identifying existing limitations and discussing possible lines for future research.

1. Introduction

In Data Science classification problems, researchers often find that they compile data with uneven class representations, which generally degrades the performance of many standard machine learning models, independently of their learning paradigms [1]. However, it is currently well known that the observed class imbalance is not the sole responsible for this undesired behaviour [2–5]. What truly hinders classification is its combination with other factors, defined in the literature as *data intrinsic characteristics* [3,5], *data difficulty factors* [4,6], or *data irregularities* [1]. These refer to several different data issues, such as class imbalance, small disjuncts, class overlap, lack of data, noisy data, dataset shift, and missing data (please refer to [5] for a comprehensive review), where class overlap has been characterised as the most harmful among them [7–9].

Note how class imbalance may not be a problem *per se*. It refers to the disproportion between class examples in the domain, which does not implicitly align with classification complexity [10]. As an example, consider a linearly separable problem, where a standard classifier will be able to obtain good performance, even if the domain is highly imbalanced. On the contrary, class overlap is undeniably problematic, even in balanced domains. It depicts a situation where examples from both classes (in binary-classification problems) are located in the same region of the data space, thus compromising the definition of clear

* Corresponding author.
E-mail address: miriams@dei.uc.pt (M.S. Santos).

https://doi.org/10.1016/j.inffus.2022.08.017
Received 16 December 2021; Received in revised form 15 June 2022; Accepted 16 August 2022
Available online 20 August 2022

decision boundaries [3,11]. In imbalanced domains, this issue is however exacerbated, since it may be in those overlapped regions that the few minority examples that exist are located. Hence, their recognition comprises a much more difficult scenario for classifiers [12,13].

Accordingly, our focus on both class imbalance and overlap is not a coincidence, since they do not have independent effects on classification performance. Nevertheless, class overlap stands as a more complex and overarching problem in classification tasks, and will therefore be given a deeper discussion throughout this work. In turn, class imbalance acts as an exacerbator and its relationship with class overlap will be depicted throughout the definition, measurement, and characterisation of the latter, notwithstanding the analysis of the synergy between both issues across several fields of Machine Learning.

The joint-effect of class imbalance and overlap has been one of the major hot topics in research for the past two decades [6,7,11,14] and is still a trending question nowadays, with applications across several fields [15–19]. Within the field of information fusion, data imperfection is one of the most challenging factors affecting fusion quality, given the complexity of application environments and associated variety and heterogeneity of data [20]. A reasonable concern regards the possibility of inadvertently creating subgroups of overlapping instances between classes during data fusion, an issue that either needs to be avoided or dealt with *a posteriori* [20]. In the line of the interest on explainability and transparency of algorithms, the use of *a posteriori* explanations, especially with respect to the generation of counterfactuals, is also deeply linked to the study of class overlap (analysing boundary or overlapping zones) [21–23]. Several applications that face these issues may be found within financial [24], medical [25–30], software [31], and network systems [32] domains, among others.

While seminal work on the topic focused on establishing class overlap as a difficulty factor for imbalanced domains, ongoing research mostly concerns the study of several forms of learning where the combination of both issues may be problematic. Accordingly, while previous work focused on artificial domains where class imbalance and overlap were synthetically generated, current research aims to characterise both problems in real-world domains.

The identification and characterisation of class overlap in imbalanced domains is, however, a subject that still troubles researchers in the field since, to this point, there is no clear, standard, well-formulated definition and measurement of class overlap for real-world domains [18]. For the most part, current research heavily relies on the *data complexity measures* originally proposed by Ho and Basu [33]. Despite the fact that many other measures have been proposed throughout the years [8,34–38], the original measures of Ho and Basu remain the most popular, promoted by open source libraries such as DCoL and ECoL [39,40].

Nevertheless, data complexity measures have the limitation of focusing on certain individual properties of data, although some data characteristics may simultaneously comprise several sources of complexity. More and more, researchers are gravitating around the idea that class overlap, especially in combination with class imbalance, is such a case [18,41,42]. It follows that class overlap arises as an heterogeneous concept, encompassing distinct representations of the problem. Accordingly, certain complexity measures are eximious in characterising some specific types of class overlap while failing to adequately capture others.

The main idea and contribution of this paper therefore consists of putting forward a unified view of the problem of class overlap in imbalanced domains, from the definition of the class overlap problem and its characterisation in all dimensions (i.e., sources of complexity), to the analysis of the most emergent topics in the field to address in the years to come. We start by introducing the idea that class overlap is currently regarded as an umbrella term that stands for a multitude of related, although distinct, problems, and discussing the key concepts associated to its definition, identification, measurement, and characterisation. Then, we map the relationship between existing data complexity measures and the specific class overlap problems they cover, proposing a new taxonomy of class overlap complexity measures. The taxonomy aggregates a comprehensive set of measures proposed over the past years, beyond the well-established data complexity measures of Ho and Basu [33]. Furthermore, this taxonomy is especially devised for the class overlap problem, while also identifying important adaptations of complexity measures that simultaneously consider the class imbalance problem. Finally, we provide a multi-view panorama on the joint-problem of class imbalance and overlap, discussing the current state of knowledge and emerging challenges across four vital areas of research in the field (Data Analysis, Data Preprocessing, Algorithm Design, and Meta-learning), and present our view on promising future directions for research in each of them.

In recent years, several outstanding survey papers have been published on the topic of learning from imbalanced datasets in the presence of data difficulty factors. A book by Fernández et al. [43] provides a comprehensive summary of the established data intrinsic characteristics and their added difficulty for classification tasks. Das et al. [1] give an impressive bird's eye view on data irregularities and their interrelation. Finally, Pattaramon et al. [18] provide an in-depth review of approaches that handle simultaneously overlapped and imbalanced domains. Similarly, the field of data complexity measures has also been a focus of intense research in the last couple of years. Most recent surveys include the research of Rivolli et al. [44], discussing existing data characterisation measures for classification datasets (including data complexity measures), and Lorena et al. [40], providing a detailed overview on data complexity measures and their use in the literature.

Contrary to previous works, this paper does not focus on presenting an exhaustive review of related work and existing approaches in the field, but rather on providing a global and unique view on the synergy between class imbalance and overlap. To the authors knowledge, this is the first work to put forward such a thorough discussion of the class overlap problem and its characterisation according to distinct representations, systematising data complexity measures towards that characterisation with the development of a new taxonomy. It also provides the most recent and comprehensive evaluation of important issues raised by the combination of class overlap and imbalance in the analysis of real-world domains.

The remaining of this paper is essentially divided into two main parts and is structured as follows. Sections 2 and 3 comprise the first half of this work and consist of a conceptual discussion of the class overlap problem. Section 2 moves towards a unifying view of the problem of class overlap, establishing the key concepts for its definition and characterisation, whereas Section 3 elaborates on our novel taxonomy of complexity measures for class overlap and illustrates the distinct representations of the problem. Then, Sections 4, 5, and 6 constitute the second half of this work, focusing on the current state of knowledge about the dual problem of class imbalance and overlap. Additionally, they are structured in a rather modular format, so that the reader may navigate them easily. Section 4 provides a panorama of the main developments across important tasks in machine learning (Data Analysis, Data Preprocessing, Algorithm Design, and Meta-learning) and the limitations they currently face. Section 5 highlights the open challenges identified within each field of Section 4 and discusses promising lines for future research. In turn, Section 6 focuses particularly on data benchmarking and open source contributions. Finally, Section 7 concludes the paper, providing an overview of the ideas discussed throughout this work and summarising important directions that the research community should debate for a renewed perspective on the problem of class overlap in imbalanced domains.

2. A unifying view on class overlap

The definition and characterisation of class imbalance is well described in the literature, where the Imbalance Ratio (IR) and the

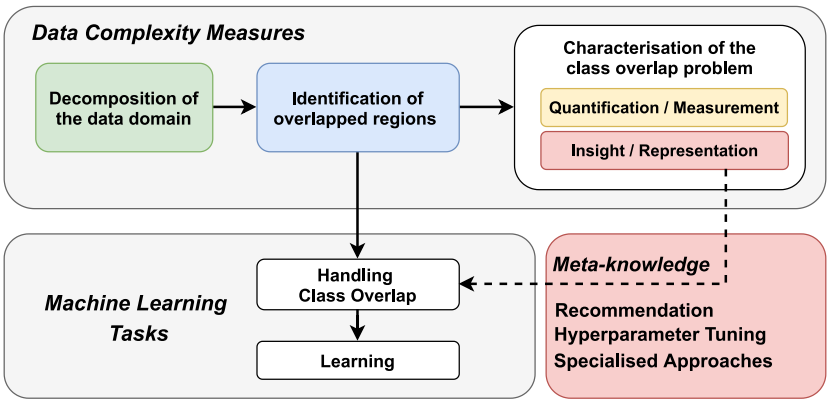

Fig. 1. An overview of the main tasks encompassed in the characterisation and analysis of the class overlap problem: (1) decomposition, (2) identification, and (3) quantification and insight. The characterisation of class overlap first requires the decomposition of the data domain into regions of interest and the identification of the problematic (overlapped) regions. Then, the chosen measure to quantify class overlap (and the insight that measure unveils) will ultimately define the representation of the problem in the domain, i.e., the specific type of class overlap that is being measured and analysed.

percentage of minority examples (%Min) constitute the standard, formal measures established in the field [43]. However, whereas class imbalance corresponds to a distribution-based irregularity, class overlap may comprise multiple sources of complexity and is therefore more complicated to assess [1,41]. Herein we provide an overview of the characterisation of class overlap, elaborating on the key concepts frequently discussed in related work, which constitutes the main contribution of this section.

The characterisation of the class overlap problem can be subdivided into three main sequential tasks, as shown in Fig. 1. First, it important to decompose the data domain into regions of interest. Then, the problematic regions (overlapped regions) need to be identified. Finally, it is possible to proceed to the quantification/measurement of class overlap, and establish its associated insight. Depending on the approaches applied to each of these tasks, class overlap may be characterised from different perspectives, leading to distinct representations of the problem (i.e., specific types of class overlap). Ultimately, each representation is associated with different measures and perceptions regarding the data domain. This measurement and characterisation of class overlap falls onto the scope of *data complexity measures* and will be addressed in Section 3. First, let us discuss the importance of establishing the key concepts and insights regarding the problem of class overlap.

Note that once the overlapping regions are identified, it is possible to handle class overlap directly (Fig. 1). This can be performed through modifications of the data domain (e.g., cleaning approaches), algorithm modification, or by handling simple and problematic regions separately, among others, depending on the end goal. However, the difference between applying ad hoc solutions that globally ease the problem and performing informed, specialised decisions based on the characteristics of the domain relies on a thoughtful understanding and characterisation of the class overlap problem. If such meta-knowledge is available, then it is possible to guide the recommendation of suitable classifiers or preprocessing techniques, the choice of suitable hyperparameters, or the design of specialised approaches. Fundamentally, determining the specific type of class overlap present in the data domain is establishing what is truly harming the machine learning tasks and, in the end, it is that insight (meta-knowledge) that guides the choice and the development of optimal solutions.

In the remainder of this section we give an overarching view of the key concepts associated with the definition of class overlap in related work, which ultimately results in the definition of distinct representations of the problem. Fig. 2 summarises both the main tasks, concepts, and insights encompassed in the characterisation of class overlap. Starting from the core of the schema, we will now move along the sequential steps required to characterise class overlap, discussing important concepts found in the literature.

Essentially, Fig. 2 corresponds to a more detailed view of the Data Complexity Measures block of Fig. 1. Accordingly, the (1) decomposition of the data domain and (2) the identification of problematic regions represent the first two tasks necessary to understand the problem of class overlap. On that note, it is important to define the concepts of Class Overlap, Overlap Regions, and Overlap Areas:

Class Overlap, Overlap Regions, and Overlap Areas: These definitions are rather intertwined since class overlap is a phenomenon that implies the existence of ambiguous regions or areas of the data space. Class overlap is often defined as (i) regions of the data space where the representation of the classes is similar [11], (ii) regions that contain a similar number of training examples from each class [3], (iii) regions with similar class priors [9] or (iv) regions containing examples from more than one class, where class boundaries overlap [5]. These definitions seek to illustrate the same idea that there may be regions of the data space that are shared by different classes. Intuitively, this complicates their discrimination, leading to a poor classification performance. Note however, how definitions (i) to (iii) refer to the concept of class overlap in a balanced scenario, equally populated by existing classes. In imbalanced domains, these definitions may not hold, as the representation of the classes in overlapping regions is not necessarily similar (nor are priors established equally for each class). A global definition of class overlap is therefore based on the existence of regions populated by examples from different classes. However, this does not prevent these regions, as well as the examples that populate them, from assuming distinct properties, leading to different representations of the problem. Accordingly, the decomposition of the data space, the identification of class overlap, and its quantification, can be performed in several ways, each focusing on different properties of the overlap regions and consequently producing different insights on the problem of class overlap. For the most part, the concept of "overlap region" is therefore a generic term, not subjected to a formal characterisation. Most often, this is also the case for "overlap area", taken as a synonym for "region", although in some related research, the overlap area is in fact defined by computing the mathematical area of overlapped regions (2-dimensional datasets) [7,18].

Once the overlap regions are identified, it is possible to move towards (3) the quantification/measurement of the class overlap problem over the domain. In that regard, related research often refers to the concept of "Overlap Degree" or "Overlap Ratio".

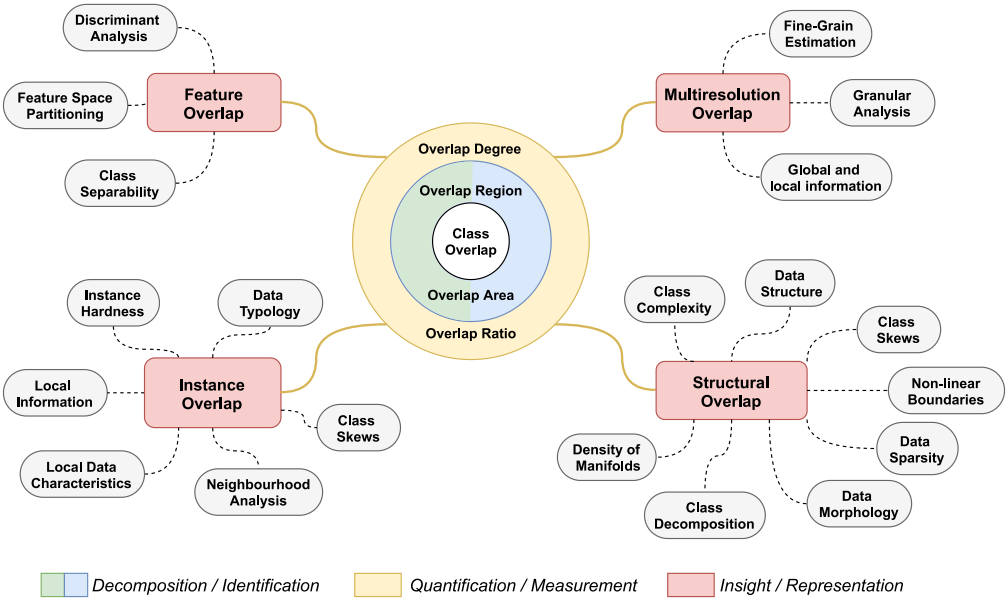

Fig. 2. An overarching view of the characterisation of the class overlap problem. Moving from the core to the peripheral parts of the schema, we may follow along the sequential steps encompassed in class overlap characterisation. First, it is necessary to (1) decompose the domains and (2) identify problematic regions (overlap regions or areas). Then it is possible to move to (3) the quantification of the class overlap problem in the domain (overlap degree or ratio). Depending on the approaches used in the previous steps, the obtained estimates will reflect distinct insights on the problem and be associated to different representations of class overlap. The established representations (Feature Overlap, Instance Overlap, Structural Overlap and Multiresolution Overlap) and associated concepts shown in the peripheral parts of the schema will be further discussed in Section 3.2.

Overlap Degree or Overlap Ratio: "Overlap Degree" is perhaps the broadest term used to describe the extent to which some domains are affected by class overlap, even when the "extent" of the problem is not mathematically defined. This occurs frequently in seminal work with synthetic data, where the overlap degree has been defined as the distance between cluster centroids of different classes [14], captured by the "extent to which adjacent regions intertwine" [11], or even not characterised numerically ([7] for atypical domains). Other seminal work estimates the overlap degree as the proportion of the domain area that is overlapped [7,45–47] (2-dimensional domains), or the proportion of examples near the decision borders [2,6,48]. In real-world domains, the quantification of class overlap is more frequent (i.e., rather than a qualitative characterisation of the problem) and is intrinsically associated to the computation of data complexity measures. In that regard, the overlap degree, sometimes referred to as "Overlap Ratio" [13,49,50], reflects a quantitative estimate of the problem of class overlap in the domain.

All in all, the concepts of overlap regions/areas and associated overlap degrees/ratios are rather generic and encompass a broad spectrum of overlap representations, depending on the strategies used to tackle the decomposition, identification and quantification of the problem. This is shown in the peripheral parts of Fig. 2 and will be clearly explained throughout the following section, where we propose a new taxonomy of class overlap measures that encompasses all three components.

3. A taxonomy of complexity measures for class overlap

Current research largely resorts to data complexity measures in order to characterise certain data characteristics. These measures are frequently organised into groups or categories, depending on the common factors each author considers in the division. By far, the most well-known grouping of complexity measures is the one defined by Ho and Basu [33], which considers three main categories: (i) measures of overlap of individual feature values, (ii) measures of separability of classes, and (iii) measures of geometry, topology, and density of manifolds. Over the years, other authors sought to complement this grouping, presenting their own division or proposing additional categories in order to characterise the prevalence of a given domain characteristic. Sotoca et al. [51] also consider three main groups of complexity measures: (i) measures of overlap, (ii) measures of class separability and (iii) measures of geometry and density. Lorena et al. [40] divide complexity measures into (i) feature-based measures, (ii) linearity measures, (iii) neighbourhood measures, (iv) network measures, (v) dimensionality measures and (vi) class imbalance measures.

For the most part, the groups discussed above do not derive from a taxonomical classification, i.e., they are defined according to each author's evaluation of common characteristics or insights among measures. The principles underlying the categorisation of measures are therefore nor explicit, nor characterised themselves. A natural consequence is that authors may include the same measure in different groups. A representative example is the grouping of F1, F2, and F3 measures, identified as *measures of overlap* in [52], as *measures of overlap of individual feature values* in [33], and as *feature-based measures* in [40]. Another example is the categorisation of T1 measure, encompassed in the *geometry, topology and density of manifolds* group in [33,53], in the *geometry and density* group in [52] and in the *neighbourhood measures* group in [40,41,54]. Throughout the years, other data complexity measures have been proposed, although they are often overlooked and included in additional categories of measures (e.g., "Other Measures" [40]).

With respect to class overlap, due to its heterogeneous nature, it is expected that several data complexity measures appear scattered across different groupings (T1 is such an example), which has several drawbacks. One is that they may not be identified as class overlap complexity measures: this is observed when measures are grouped based on the object of analysis (e.g., feature-based measures, neighbourhood measures), rather than according to the insight they provide over the domain (e.g., feature overlap, instance overlap). Other is that some recent measures that characterise class overlap are either described as general complexity measures, included in a separate category (e.g., "Other Measures"), or do not figure among well-established groupings. Finally, some of the existing groups may be misleading

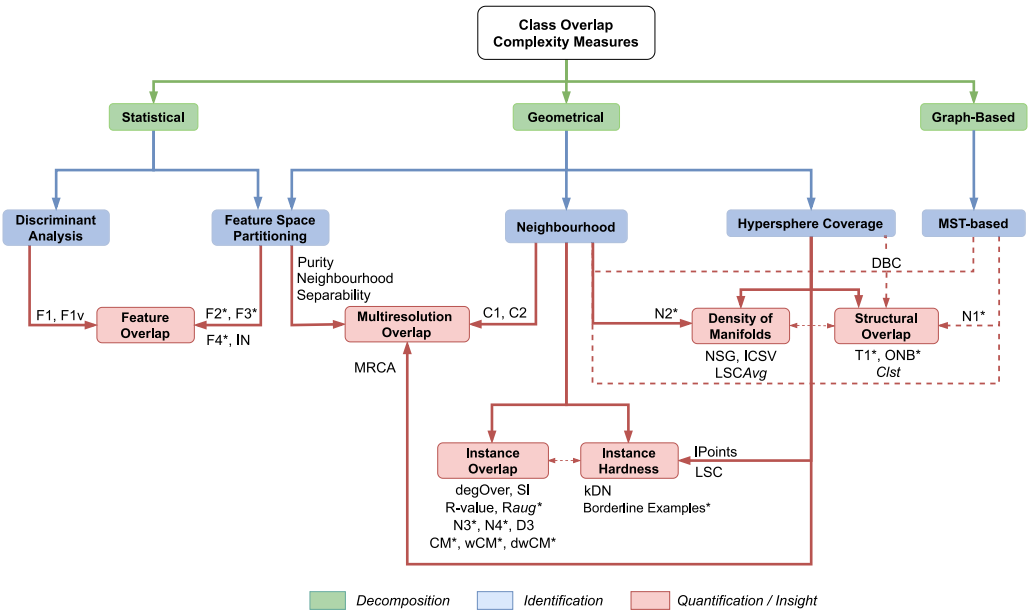

Fig. 3. Taxonomy of class overlap complexity measures. Different groups can be established depending on the level of the analysis. In the tree structure, class overlap measures are divided in what concerns their approach to decompose the data domain, identify regions of interest and quantify class overlap. Measures marked with an asterisk are those for which adaptations to imbalanced domains have been explored in the literature.

by defining categories of *measures of overlap* that comprise only measures that capture only one specific type of class overlap (e.g., feature overlap).

We advocate that data complexity measures should be grouped according to the insight they provide over the domain, and in the particular case of class overlap, that a taxonomy of complexity measures should attend to its heterogeneous nature. It would therefore be instrumental to define a taxonomy of class overlap measures that attends to its different representations and sources of complexity. However, no such characterisation currently exists. To put forward such taxonomy is the main contribution of this section.

As previously discussed, the characterisation of class overlap is intrinsically tied to the definition and quantification of problematic regions in data. Accordingly, along this section, we devise a taxonomy of complexity measures for class overlap based on the strategies used to address the three main identified components of overlap characterisation: (1) decomposition of the data space, (2) identification of problematic regions, and (3) quantification and insight of the overlap problem in the domain.

The proposed taxonomy is presented as a tree structure (Fig. 3), based on the sequential tasks of Figs. 1 and 2. Class overlap measures are first divided depending on their decomposition of the data space. As we move down each path, further groups arise, depending on the identification of problematic areas and ultimately, on the class overlap representations they are able to capture.

Rather than focusing solely on the well-known measures of Ho and Basu [33], we consider a larger set of measures proposed throughout the years. The relationship between measures is also characterised, since some measures based on different paradigms may provide similar insights, whereas others are complementary. Complexity measures that have been previously studied in imbalanced contexts are also identified. The reader may find additional information regarding the mentioned complexity measures in [19]. Additionally, to support the reading of this section, the characteristics of each complexity measure were summarised in Table 1. In the remainder of this section we will elaborate on further aspects of the proposed taxonomy. First, we start by defining and describing the essential components of class overlap characterisation (Section 3.1). We mainly focus on components (1) and (2), whereas (3), comprising the final proposed representations of class overlap and respective insights, is further discussed on Section 3.2, alongside their associated complexity measures. Finally, we end this section with an evaluation of the proposed taxonomy, as well as its implications regarding future research (Section 3.3).

3.1. Components for defining a taxonomy of class overlap measures

Essentially, all overlap measures require three components:

1. **A component to decompose the data domain into regions of interest:** We consider three main approaches to divide the feature space into regions of interest. Although all are distance-based, they rely on different types of distances:
   - **Statistical Distance:** Based on the distance between class distributions (e.g., Fisher Linear Discriminant);
   - **Geometrical Distance:** Based on the distance between pairs of data examples (e.g., Euclidean Distance);
   - **Graph-Based Distance:** Based on the geodesic distance (e.g., Minimum Spanning Trees).
2. **A component to identify problematic regions of interest.** We consider the following strategies for the identification of problematic regions:
   - **Discriminant Analysis:** The properties of class distributions are analysed in order to determine the discriminative power of features. Problematic regions are those where classes remain overlapped in the projections with maximum separability;
   - **Feature Space Partitioning:** The feature space is partitioned into certain ranges or into a specified number of intervals where the properties of data are then analysed. Problematic regions are delimited in specific ranges of the feature space;
   - **Neighbourhood Analysis:** The data domain is analysed at a local level, based on the neighbourhood characteristics of examples. Problematic regions are those associated to larger errors of the k-nearest neighbour classifier;

Table 1
Main characteristics of class overlap complexity measures. For each complexity measure, it is identified which class overlap representation it is able to capture, "Representation"; its abbreviation and name, "Abbr." and "Measure"; its complexity interpretation, "Complexity" ("++" denotes that higher values of the measure indicate more overlapped domains, whereas "−−" denotes that lower values indicate more overlapped domains, according to the formulation established in Santos et al. [19]); its taxonomical classification, "Taxonomy"; and whether it has been previously investigated in imbalanced domains, "Imbalanced Data" (C.D.: Class Decomposition, IR: Imbalance Ratio).

Representation	Abbr.	Measure	Complexity	Characteristics	Taxonomy	Imbalanced data
Feature Overlap	F1	Maximum Fisher's Discriminant Ratio	++	Determines the maximum discriminative power of features.	*Statistical: Discriminant Analysis* ➡ *Feature Overlap*	No
	F1v	Directional Vector Maximum Fisher's Discriminant Ratio	++	Determines the data projection with maximum separability.	*Statistical: Discriminant Analysis* ➡ *Feature Overlap*	No
	F2	Volume of Overlapping Region	++	Measures the volume of the overlapping region by determining the overlap range of each feature.	*Statistical: Feature Space Partitioning* ➡ *Feature Overlap*	Yes (C.D.)
	F3	Maximum Individual Feature Efficiency	++	Determines the minimum amount of overlap between feature values of different classes.	*Statistical: Feature Space Partitioning* ➡ *Feature Overlap*	Yes (C.D.)
	F4	Collective Feature Efficiency	++	Returns the ratio of examples that could not be separated considering the efficiency of all features.	*Statistical: Feature Space Partitioning* ➡ *Feature Overlap*	Yes (C.D.)
	IN	Input Noise	++	Determines the amount of overlap across all dimensions in data.	*Statistical: Feature Space Partitioning* ➡ *Feature Overlap*	No
Instance Overlap	R-value	R-value	++	Measures the degree of overlap between two classes by determining the number of points of each class that fall onto overlap regions between classes.	*Geometrical: Neighbourhood* ➡ *Instance Overlap*	No
	Raug	Augmented R-value	++	Extends R-value taking the Imbalance Ratio into account.	*Geometrical: Neighbourhood* ➡ *Instance Overlap*	Yes (IR)
	$degOver$	Degree of Overlap	++	Determines the ratio of overlapping examples in data by considering conflicting class neighbourhoods.	*Geometrical: Neighbourhood* ➡ *Instance Overlap*	No
	N3	Error Rate of the Nearest Neighbour Classifier	++	Measures the error rate of the Nearest Neighbour classifier (1NN), estimated using a Leave-One-Out cross-validation.	*Geometrical: Neighbourhood* ➡ *Instance Overlap*	Yes (C.D)
	SI	Separability Index	++	Determines the proportion of points whose class is the same as of its nearest neighbour.	*Geometrical: Neighbourhood* ➡ *Instance Overlap*	No
	N4	Non-linearity of the Nearest Neighbour Classifier	++	Measures the 1NN error on a set of new synthetic examples generated by interpolating pairs of data examples from the same class, chosen randomly.	*Geometrical: Neighbourhood* ➡ *Instance Overlap*	Yes (C.D.)
	kDN	k-Disagreeing Neighbours	++	For each data example, kDN measures the percentage of its k nearest neighbours that do not share its class.	*Geometrical: Neighbourhood* ➡ *Instance Hardness*	No
	D3	Class Density in the Overlap Region	++	Determines, for each class, the number of examples that lie in regions populated by a different class.	*Geometrical: Neighbourhood* ➡ *Instance Overlap*	No
	CM wCM dwCM	Complexity Metric based on k-Nearest Neighbours	++	Measures the proportion of difficult examples in data, considering conflicting class neighbourhoods.	*Geometrical: Neighbourhood* ➡ *Instance Overlap*	Yes (C.D.)
	Borderline Examples	Borderline Examples	++	Determines the percentage of borderline examples in data, according to a data typology that divides examples into safe, borderline, rare and outlier categories.	*Geometrical: Neighbourhood* ➡ *Instance Hardness*	Yes (C.D.)
	IPoints	Number of Invasive Points	++	Finds the number of invasive points in data, by analysing the local set of each data example.	*Geometrical: Hypersphere Coverage* ➡ *Instance Hardness*	No
Structural Overlap	N1	Fraction of Borderline Points	++	Measures the proportion of examples that are connected to the opposite class by an edge in a Minimum Spanning Tree.	*Graph-Based: MST-based* ➡ *Structural Overlap* *Geometrical: Neighbourhood* ➡ *Structural Overlap*	Yes (C.D.)
	T1	Fraction of Hyperspheres Covering Data	++	Determines the number of hyperspheres of the same class necessary to cover the entire data domain.	*Geometrical: Hypersphere Coverage* ➡ *Structural Overlap*	Yes (C.D.)
	LSCAvg	Local Set Average Cardinality	−−	Determines the average local set cardinality considering all points in data.	*Geometrical: Hypersphere Coverage* ➡ *Density of Manifolds*	No
	$Clst$	Number of Clusters	++	Determines the number of clusters of the same class necessary to cover the data domain, performing the clustering procedure according to the local set cardinality of data examples.	*Geometrical: Hypersphere Coverage* ➡ *Structural Overlap*	No
	ONB	Overlap Number of Balls	++	Determines the number of balls of the same class necessary to cover the data space using the Pure Class Cover Catch Digraph to determine the maximum radii of all examples in data.	*Geometrical: Hypersphere Coverage* ➡ *Structural Overlap*	Yes (C.D.)
	DBC	Decision Boundary Complexity	++	Determines the interleaving of hyperspheres of different classes, by determining the number of connected centres of different classes in a MST built with the final hyperspheres determined with the T1 measure.	*Geometrical: Hypersphere Coverage* ➡ *Structural Overlap* *Graph-Based: MST-based* ➡ *Structural Overlap* *Geometrical: Neighbourhood* ➡ *Structural Overlap*	No
	N2	Ratio of intra/extra Class Nearest Neighbour Distance	++	Illustrates a trade-off between the intra-class distances and inter-class distances.	*Geometrical: Neighbourhood* ➡ *Density of Manifolds*	Yes (C.D.)
	NSG	Number of Samples per Group	−−	Determines the average number of examples in each hypersphere found with the T1 measure.	*Geometrical: Hypersphere Coverage* ➡ *Density of Manifolds*	No
	ICSV	Inter-class Scale Variation	++	Measures the standard deviation of the densities of the hyperspheres found with the T1 measure.	*Geometrical: Hypersphere Coverage* ➡ *Density of Manifolds*	No
Multiresolution Overlap	MRCA	Multiresolution Complexity Analysis	++	Identifies regions of different complexity in the domain by profiling data examples according to the characteristics of their surrounding hyperspheres, constructed with increasing radii.	*Geometrical: Hypersphere Coverage* ➡ *Multiresolution Overlap*	No
	C1	Case Base Complexity Profile	++	Determines the average complexity of the domain by analysing the complexity profile of each data example at increasing neighbourhood sizes.	*Geometrical: Neighbourhood* ➡ *Multiresolution Overlap*	No
	C2	Similarity-Weighted Case Base Complexity Profile	++	Modification of C1 considering weighted contributions of nearest neighbours.	*Geometrical: Neighbourhood* ➡ *Multiresolution Overlap*	No
	Purity	Purity	−−	Estimates the purity of the domain by focusing on the class distribution of recursive partitions of the data space (cells) defined at several resolutions.	*Geometrical: Feature Space Partitioning* ➡ *Multiresolution Overlap*	No
	Neighbourhood Separability	Neighbourhood Separability	−−	Estimates the separability of the domain by focusing on the neighbourhood characteristics of each example comprised inside cells defined at several resolutions.	*Geometrical: Feature Space Partitioning* ➡ *Multiresolution Overlap*	No

- **Hypersphere Coverage:** The necessary number of subsets (hyperspheres) to cover the entire domain is found. Problematic regions are those encompassed in hyperspheres with smaller radii;
- **Minimum Spanning Trees:** The data domain is represented by a graph (often a minimum spanning tree). Problematic regions are identified by directly connected vertices with disagreeing class memberships.

3. **A component for quantifying the overlap problem in the problematic areas of interest.** This component returns the final groups of the tree structure, consisting in the ultimate division between overlap measures. For that reason, we will discuss each group in detail throughout the following sections, along with the measures they include and the insights they provide.

By addressing the definition and quantification of problematic regions differently, complexity measures characterise class overlap from different perspectives. Indeed, in real-world domains, problematic regions often present certain properties that have an impact on the definition and measurement of class overlap (e.g., class imbalance, local imbalance, class decomposition, non-linear boundaries, different types of examples in data) [2,6,7,15]. These characteristics of data may therefore give rise to different representations of class overlap, and certain measures may successfully characterise some, while failing to uncover others. The final groups of the proposed taxonomy associate the complexity measures to the representations of class overlap they intend to characterise, and are thoroughly described in what follows.

3.2. Representations of class overlap

Formally, we recognise four main representations (i.e., specific types) of class overlap: Feature Overlap, Instance Overlap, Structural Overlap, and Multiresolution Overlap. There are however some subgroups that somewhat complement the characterisation of certain representations (Instance Hardness and Density of Manifolds). They will be discussed within the respective groups (Instance Overlap and Structural Overlap, respectively).

3.2.1. Feature overlap

Class overlap is often referred to as "class separability" [5,9,55]. This term refers to the degree to which classes may be separated by discriminative rules, i.e., the degree to which good decision boundaries may be found. Hence, it provides an interpretation of class overlap via its contrary, i.e., an overlapped domain is one where the class separability is low.

Feature Overlap measures are intrinsically associated with the concept of class separability, i.e., they aim to characterise the discriminative power of features in data or, accordingly, the class overlap of individual features in data. Some measures estimate class overlap by looking for the most discriminative projections of data (F1, F1v) [33,40], where others resort to feature space partitioning to delimit overlap regions, based on the properties of class distributions (F2, F3, F4, IN) [33,39,56].

By focusing on the individual properties of features, these measures may fail to capture other idiosyncrasies of class overlap. Take for instance the scenario depicted in Fig. 4. F1 measures the highest discriminative power for all features in data, i.e., it returns the minimum overlap of individual features found in the domain. Accordingly, the scenarios in Fig. 4 reveal the same discriminative power: feature f_1 has the same (and highest) F1 value in both cases. However, the individual overlap in feature f_2 is different, which makes these scenarios different in terms of classification difficulty (as emphasised by the superimposed optimal linear discriminant). In turn, marked points illustrate the facet of the problem measured by Instance Overlap. Rather than analysing feature separability, instance overlap – described in the following section – captures the amount of conflicting examples in data through the analysis of their neighbourhood, thus obtaining different estimates for the presented scenarios.

Other limitations of feature overlap measures have already been described in the literature [35,40]. First, these measures presuppose their application over continuous features. Then, with the exception of F1v, they assume that the decision boundary between classes is perpendicular to one of the feature axis. Measures based on feature space partitioning (F2, F3, F4, IN) are additionally susceptible to disjunct concepts (a situation where features present more than one valid interval), and noisy data.

3.2.2. Instance overlap

Instance Overlap measures are deeply linked to the exploration of "local data characteristics" [57] and comprise a local, rather than a global, characterisation of domains. These characteristics are often approximated by analysing the neighbourhood of data examples and determining their complexity accordingly. This "complexity" is often associated to the error of the k-Nearest Neighbour (kNN) classifier and is used to characterise class overlap by focusing on the amount of overlapped examples in data, i.e., those that are misclassified by kNN. Instance Overlap measures include R-value [58], R_{aug} [38], *degOver* [15], N3 [33], SI [59,60], D3 [52], N4 [33], CM, wCM, and dwCM [17,34], which provide an overall insight on the amount of overlapped examples in the entire domain, and kDN [8], Borderline Examples [2], IPoints and LSC [36], which, despite providing similar insights, are more aligned with the idea of estimating the complexity of individual examples in data, associated to the concepts of "instance hardness" [8] and "data typology" [2].

"Instance Hardness" and "Data Typology" reflect the idea that not all examples in data are equal for classification tasks. On the contrary, depending on the local characterisation of class distributions, some examples may be harder to learn than others. "Instance Hardness" corresponds to the likelihood of an example to be misclassified, for which class overlap is the principal contributor [8]. In turn, "Data Typology" comprehends the division of data examples according to four types: *safe*, *borderline*, *rare*, and *outlier* examples [61]. Note that ultimately, the typology of examples depends on the endgame and desired treatment of different types of examples, and therefore it is not uncommon to find other notions of *redundant*, *noisy*, *danger*, or *unsafe* examples [10,62,63]. Overall, since *borderline* examples are those located in the borderline between classes, where their discrimination becomes complicated, they are highly associated with the definition of class overlap [2,6,48,61]. Nevertheless, it may also be important to consider overlapped examples scattered across the entire domain, i.e., those that, although farther from the border, also contribute to class overlap [64]. In that sense, borderline examples are considered a subset of overlapped examples, and class overlap measures may either consider solely the borderline regions between classes or the entire domain. This ultimately relies on each measure's setting regarding the size of local neighbourhoods (k value) and/or the tolerance threshold which distinguishes an overlapped from a non-overlapped example.

The concept of "Class Distribution Skew" is also worthy of discussion within the problem of class overlap [1,7]. In addition to situations where classes are intertwined, class overlap may possess other structural biases, where one class is dominant in the overlap region. Such a phenomenon may arise due to the presence of local imbalance in the overlap region, or irrespective of class imbalance, e.g., due to differences in class densities (one class is sparse in the overlap region whereas the other is dense). Some authors refer to this phenomenon as "local densities" [7], while other describe it as a distribution skew or "class skew" [1]. In such scenarios, instance overlap measures, due to their flexibility (variable neighbourhood definition), may be helpful in capturing the degradation caused by class overlap.

Nevertheless, instance overlap measures, focusing on the properties of individual examples in data, disregard the characterisation of overlap regions themselves. In general, instance overlap measures

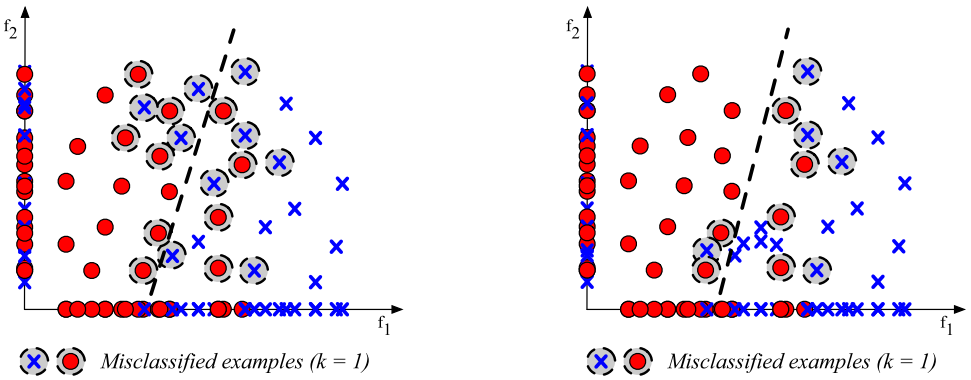

Fig. 4. Example of F1 computation for two domains, where data examples are projected onto the axis. The F1 measure outputs the same value of class overlap for both domains, despite the fact that the problem affects domains differently, as indicated by the superimposed optimal linear discriminant. Note how $f1$ has the same and maximum discriminative power in both domains, whereas the individual overlap in $f2$ is different between domains. F1 therefore captures one facet of class overlap (feature overlap) but it may not provide a full characterisation of the class overlap problem. As an example, marked points illustrate a representation of instance overlap, identifying data points which are misclassified by their nearest neighbour ($k = 1$). Different estimates of class overlap are obtained for each domain, namely 19/35 = 54.3% and 11/35 = 31.4% for the left-side and right-side, respectively.

are concerned with the class membership of examples within a k-neighbourhood, regardless of the actual distance between them. It follows that, given two examples that are each other's nearest neighbours, instance overlap measures cannot distinguish a situation where they share similar values in the feature space from a situation where they have rather different feature values. Ultimately, despite being each other's closest neighbours, the examples may belong to distinct regions of the data space where there is no class overlap. Similarly, in the borderline between classes, instance overlap measures may also produce erroneous estimates of class overlap in some scenarios.

Consider Fig. 5a, where the distance between examples on class boundaries is smaller than the distance between examples of the same class. Instance overlap measures, focusing on local properties of data, will produce biased class overlap estimates even though the domain illustrates a linearly separable problem. Additionally, domains where the properties of examples are the same at a local level may be indistinguishable. Consider Figs. 5a and c, which comprise examples with similar local neighbourhoods. Oblivious to the global properties of problematic regions, instance overlap measures will output similar values of class overlap for both domains. In turn, note how analysing the global properties (e.g., structure) of problematic regions (Figs. 5b and d) provides a different insight on the characterisation of the class overlap problem of Figs. 5a and c.

Increasing the value of the k is one way to move towards a more global view of the domain [7,34]. Note how the scenario depicted in Fig. 5a would be distinguishable from (c) if instead of $k = 1$, we were to consider $k = 3$ or 5: in (c), we would find a larger number of examples with conflicting class neighbourhoods. However, optimal values of k are hard to determine, especially in the presence of domain peculiarities such as class skews: k values that correctly characterise one region may produce biased estimates in another.

Similarly, categorising examples into several types is a way of approximating the global properties of data, which provides additional insight on the domain; yet it is still based on a local analysis paradigm (dependent on the k hyperparameter configuration). These are intrinsic limitations of instance or neighbourhood-based identification and may be attenuated by a characterisation of problematic regions themselves, focusing on a global analysis of the domain.

As an example, consider Fig. 6, which characterises two data domains (a and d) from a local to a global perspective. Note how (a) and (d) return the same overlap value ($k = 1$), despite depicting different representations of class overlap. The identification of different types of examples ($k = 5$, b and e) reveals that the domains are indeed conceptually different: a/b observe a more classical class overlap (complicated borderline regions), whereas d/e depict a situation where complicated examples from one class (blue crosses appearing as rare and outlier examples) are scattered throughout regions of the other. The characterisation of the class overlap problem in each domain may be complemented by focusing on global, structural properties of data: (c) characterises the domain as having two well-defined concepts and a confounding boundary (balls of both classes with smaller radii, containing only one example and close to each other), whereas (f) identifies a well-defined region of one class (blue crosses comprised in a lower number of balls with large radii and local sets) and another region with higher class decomposition (red points comprised in a larger number of balls with variable local sets) contaminated with scattered examples of the opposite class (blue crosses in balls of smaller radii, containing only one example, close to larger balls of the other class, with higher local sets).

3.2.3. Structural overlap

Recognised as the most impactful issue for prediction tasks [7,9], class overlap is also often used interchangeably with the term "class complexity" [55]. We have seen this for instance overlap measures, where class overlap is associated with the complexity of individual examples in data and often evaluated on the basis of disagreeing neighbourhoods of examples (overlapped or "complex" examples) [17, 34]. Beyond this, recall that class overlap aggregates a multitude of complexity sources, as we have been discussing so far. In particular, data morphology (data topology, shape or structure) may have hidden dependencies on the problem. On the one hand, the global characteristics of the domain (e.g., class decomposition, complexity of the decision boundaries, data sparsity) influence the identification of problematic regions and consequently the quantification and characterisation of class overlap. On the other hand, class overlap directly affects the shape of the decision boundaries between classes and may create additional complications such class skews, changing the structural properties of the domains. In fact, recent research is gravitating towards the idea that complexity measures related to data morphology may prove good predictors of class overlap, especially in the context of imbalanced domains [41,42].

Structural Overlap measures are more attentive to the internal structure of classes (data morphology) when evaluating problematic regions. Some measures analyse the properties of a minimum spanning tree (MST) built over the data domain to identify complicated regions where classes intertwine (N1 [33]). Others approach the identification of class overlap using the notion of hypersphere coverage, where the domain is entirely divided into subsets comprising only examples of the same class (T1 [33], *Clst* [36], ONB [41]). Some consider both MST and hypersphere coverage (DBC [65]). Additionally, we refer to

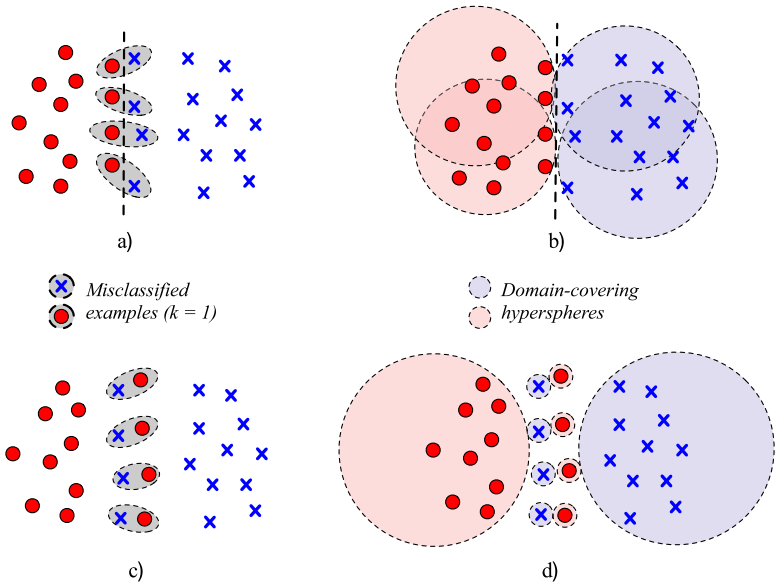

Fig. 5. Comparing local (a and c) versus global (b and d) information. Focusing on local information, instance overlap measures may not be able to capture certain properties of the domains that affect class overlap. Note how (a) and (c) result in similar class overlap characterisations (same percentage of conflicting examples), despite the fact that (a) is linearly separable, as indicated by the superimposed linear discriminant. Analysing the structure of problematic regions (b and d) provides different insights on the characterisation of the class overlap problem, where (d) requires a higher number of hyperspheres to cover the entire domain, thus illustrating a more intertwined scenario.

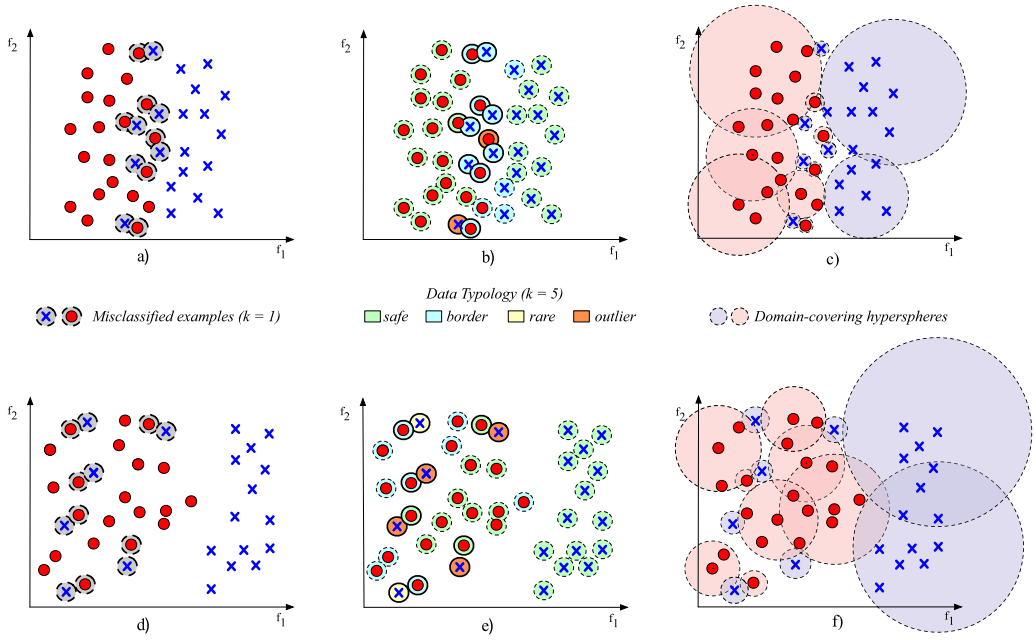

Fig. 6. Characterisation of two domains affected by class overlap, moving from a local (a and d) to a global analysis (c and f): in scenarios (a) and (d), class overlap is estimated through the number of conflicting examples (nearest enemies); in (b) and (e) the data typology of the domains is used to characterise class overlap via borderline or non-safe examples; in (c) and (f) the number of hyperspheres needed to cover the domains is computed to characterise how intertwined the domains are. Instance overlap measures define class overlap by analysing the properties of individual examples, thus neglecting certain structural characteristics of the domain: note how (a) and (d) return the same percentage of complicated instances, despite depicting different representations of class overlap. Studying the data typology is a way of approximating the global properties of the domain, combining both local and global information (although still dependent on k hyperparameter configuration): the data typology reveals that a/b illustrate complicated borderline regions, whereas d/e depict a scenario where examples of one class are scattered throughout regions of the other. The characterisation of the class overlap problem may be complemented by structural overlap measures, focusing on global, rather than local, characteristics of the domains: note how (c) illustrates two well-defined concepts with a complicated decision boundary, while (f) shows a well-defined region of one class with some instances contaminating the region occupied by the other class.

a subset of structural overlap measures ("Density of Manifolds" group) that complements the characterisation of class overlap by adding local information to data morphology, i.e., focusing on data density/sparsity. These measures characterise the average number and dispersion of examples comprised within the hyperspheres that cover the domain (NSG and ICSV [56]), describe the within- and between-class spread (N2 [33]), or the average local set cardinality of examples in the domain (LSC$_{Avg}$ [36]).

Recall the domains of Fig. 6, where the analysis of global, structural information (Figs. 6c and f) supports the distinction between a domain with complicated borderline regions (Fig. 6a) and a domain with a large amount of intrusive points (Fig. 6d). Figs. 6c and 6f are in fact representative of structural overlap and illustrate the computation of

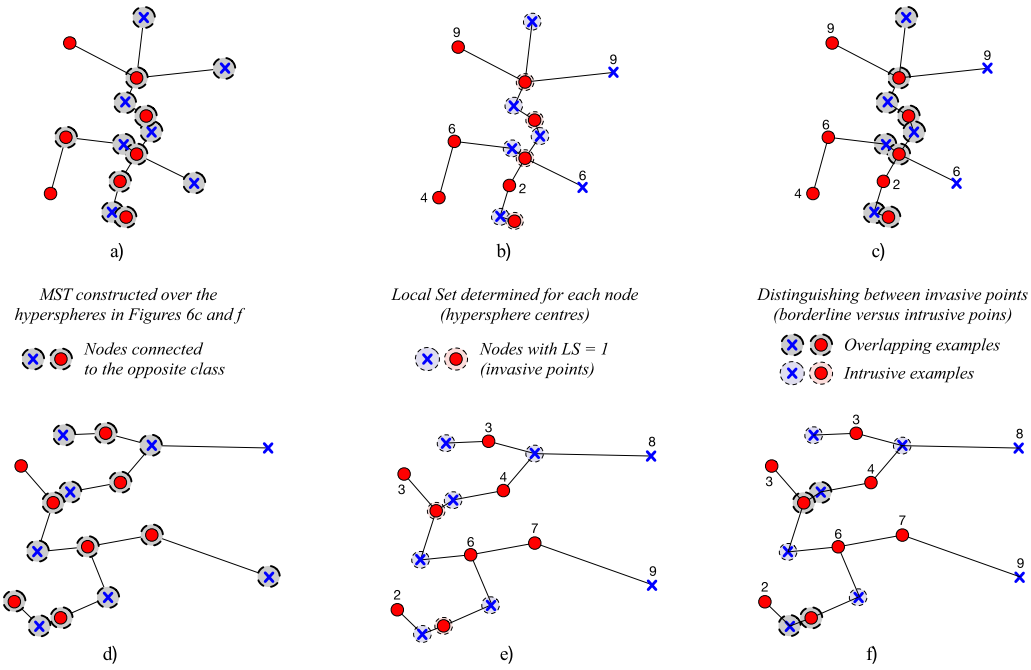

Fig. 7. Exploring the structural properties of the domain may be fundamental to derive a more accurate characterisation of class overlap. Starting with the domains of Figs. 6c and f, the scenarios in (a) and (d) assess the interleaving of classes along the decision boundary of each domain, by building a MST considering the hyperspheres' centres and determining the number of connected nodes, i.e., computing DBC. Nevertheless, note how complexity measures that focus on individual characteristics of data, such as DBC in (a) and (d), may not extract perceptive insights. In this regard, exploring additional information on the domain, such as considering the local set of each node (i.e., each hypersphere centre) as represented in (b) and (e), may lead to a better understanding of what is truly harming the domains, identifying invasive points. By combining information regarding invasive points with the structure of the MST solution, it is possible to distinguish between domains comprising mostly borderline examples, such as in (c), and intrusive examples, such as in (f), enabling the development of specialised solutions for each scenario.

Clst [36], which divides the data domain into clusters of the same class. However, despite the fact that the domains are easily distinguished when visualised, their *Clst* values are rather similar, since *Clst* is only concerned with the number of total clusters in data, regardless of their radius, their local sets (how many examples they cover), or the distance between them.

A way to enhance this characterisation would be to analyse additional structural information, such as assessing the interleaving of classes along the decision boundary of each domain. Accordingly, Figs. 7a and 7d illustrate a representation of DBC [65], which creates a MST using the cluster centres defined by *Clst* and determines the number of connected centres of different classes. As in the previous case, although the problem of class overlap is conceptually different when assessed visually, DBC also returns similar values, since the number of connected nodes of opposite classes is similar for both domains. The analysis of NSG [56], which returns the average size of clusters, would yield identical conclusions to those of the previous measures.

Note how the difficulty in distinguishing the domains via existing complexity measures is due to their focus on individual properties of data: *Clst* and NSG disregard the characterisation of clusters whereas DBC neglects other properties of the MST (e.g., edge weights, local sets of connected nodes). Alternatively, Figs. 7b and e characterise the domains by combining several structural overlap measures. Accordingly, they incorporate information regarding class decomposition (starting with the solution defined by *Clst*), complexity of decision boundaries (considering the solution achieved by DBC), and density of manifolds (considering the local set cardinality of each node in the MST).[1] Contrarily to Figs. 7a and d, the marked points represent clusters that include only one example (the core) and whose local set contains only the core itself, defined as "invasive points"(IPoints) [36]. Now, despite the number of invasive points is similar in both domains, it is possible to differentiate (i) situations where these points are "strongly connected"[2] to others of the same type of the opposite class, identifying examples located in overlapping regions of the data space, from (ii) situations where these points are connected to nodes of the opposite class with larger local sets, identifying examples that somewhat infiltrate the other class. Hence, Fig. 7c illustrates a domain where all of its invasive points strongly connect to others of the same type (and of the opposite class), suggesting that class overlap is the main complexity factor affecting the domain (9 out of 15 nodes represent complicated borderline regions, which amounts to a class overlap of 60%), caused by overlapping class borders. In turn, Fig. 7f reveals that only 4 out of 16 nodes (25%) are responsible for class overlap (4 invasive points strongly connected), whereas the remaining 4 identified points are intruding the opposite class, and may indicate different issues: either

[1] Note how despite the fact that LSC*Avg* is comprised in the Structural Overlap group (as it estimates the density of manifolds in the domain), and that LSC and IPoints derive from structural information (i.e., hypersphere coverage), they can be used to add local information regarding the internal class structures found in data. In fact, LSC may be an indicator of instance hardness and instance overlap, identifying examples whose local set cardinality is low.

[2] Note that our purpose is not to derive a new complexity measure for class overlap. With this example, we explore the investigation of additional properties of the MST (namely edge weights) as well as density and local information (local set cardinality) to complement the characterisation of class overlap. Combining distinct sources of information allows to distinguish shorter, stronger connections between nodes, from weaker connections, where edges between nodes are longer. To determine whether an invasive example is responsible for class overlap or is infiltrating the opposite class – in the case that an invasive point is connected to both an invasive point and other nodes of higher cardinality (all of the opposite class) – it is possible to adjust the edge weights by the local set cardinality of connected nodes (e.g., $w_i = \frac{1}{d_i} \times LSC_{node_i}$). Nevertheless, the main purpose of the example remains to highlight the advantage of considering multiple sources of complexity when characterising class overlap.

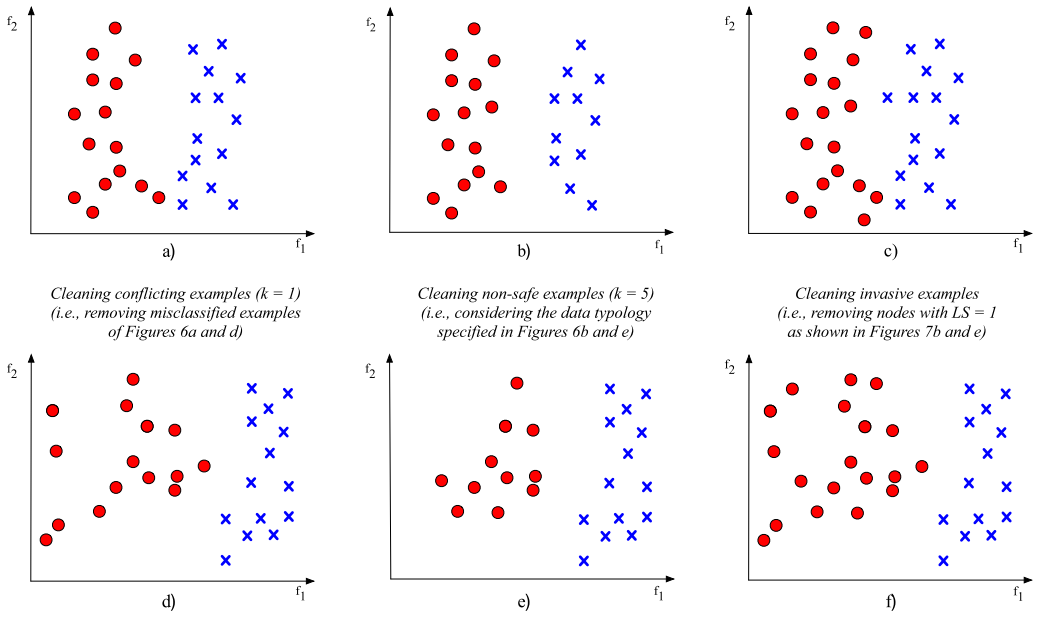

Fig. 8. Impact of considering structural information in the characterisation of class overlap. Scenarios (a) and (d) illustrate the solution achieved by removing all conflicting examples according to Figs. 6a and d (examples misclassified by their nearest neighbour ($k = 1$) are eliminated). In (b) and (e), all examples that do not belong to the "safe" category are removed (i.e., all the borderline, rare, and outlier examples), following the data typology of Figs. 6b and e. Finally, (c) and (f) illustrate the removal of the invasive points shown in Figs. 7b and e.

representing noisy data [5], or suggesting the existence of valid, though underrepresented, sub-concepts in data (a situation likely to arise in the case of imbalanced data [61]).

Let us end this discussion by analysing the impact of considering structural information in the characterisation of class overlap. Fig. 8 shows different cleaning solutions for the original domains of Figs. 6a and d (top and bottom rows of Fig. 8, respectively).

Despite the fact that all characterisations of class overlap lead to solutions with simplified, clear decision boundaries, i.e., eliminating the problem of class overlap, they differ in what concerns both the amount of cleaning performed and the ability to retain the original structure of data. Approaches relying solely on instance overlap (Figs. 8a, b, d, and e) tend to be more conservative when compared to those that incorporate structural information (Figs. 8c and f). Also, note that since Figs. 6b and e consider more global information on the data domain than 6a and d (via data typology), the former solutions are more conservative than the latter. This is due to (i) the larger neighbourhood considered: $k = 5$ versus $k = 1$, identifying only nearest-enemies (please refer to Fig. 6b where more examples are considered conflicting), and (ii) the *borderline* category often assigned to examples in the neighbourhood of rare and outlier examples, which may not represent valid class concepts, but rather intrusive/noisy points, affecting mainly domain 6e.[3]

In turn, solutions 8c and f are the less invasive, i.e., the class overlap problem is solved while removing a smaller amount of examples and retaining most of the original internal structure of data. Finally, note how for domains with less complex data structure/morphology, instance overlap measures are able to accurately characterise the problem of class overlap, whereas structural information needs to be considered when dealing with domains presenting additional sources of complexity. On that note, although we may argue that structural overlap measures focus on data characteristics unrelated to class overlap, in the sense that they describe other general properties of the domains (e.g., geometry, topology, density), we advocate that class overlap cannot be fully understood irrespective of structural information, since the global properties of the domains affect its identification and characterisation.

3.2.4. Multiresolution overlap measures

Multiresolution Overlap measures characterise class overlap by providing a trade-off between global and local data characteristics (Fig. 9). Some are more closely related to the previous ideas of using hyperspheres (MRCA [37]) or k-neighbourhoods (C1 and C2 [35,66]) to define regions of the space where class overlap can be analysed. Others are associated with feature space partitioning, where features are divided into several intervals to assess the properties of class overlap (Purity and Neighbourhood Separability [67,68]). Nevertheless, the main idea than binds these measures together is that they operate by moving iteratively from a global to a local analysis of the domains (fine-grain search criteria). They recursively define hyperspheres, neighbourhoods, or feature partitions at different resolutions, all of which are analysed characterise to the problem of class overlap, combining both structural and local information.

3.3. Evaluation of the proposed taxonomy

Along the previous sections, we have been discussing the idea that, in real-world domains, class overlap often aggregates information on different data characteristics, and therefore it is important to establish the insight that different complexity measures provide to fully characterise the problem. To standardise existing types of class overlap, we established a novel taxonomy that defines four main groups of class overlap representations and associated complexity measures, while describing their perception on the class overlap problem as well as their intrinsic limitations. In this section, we discuss some further details of the proposed taxonomy, and elaborate on its implications for future research in the field.

[3] Note that in imbalanced domains, there a difference between *rare* and *outlier* examples, and noisy data (please refer to [61]), given that distant, isolated minority examples may result from an insufficient representation of the minority class in certain regions of the data space. Accordingly, *rare* and *outlier* examples may represent valid sub-concepts rather than noise. Nevertheless, the given example (Fig. 6e), represents a balanced domain where rare and outlier points are not distant or isolated examples, but rather infiltrating the opposite class and do not constitute interesting class concepts.

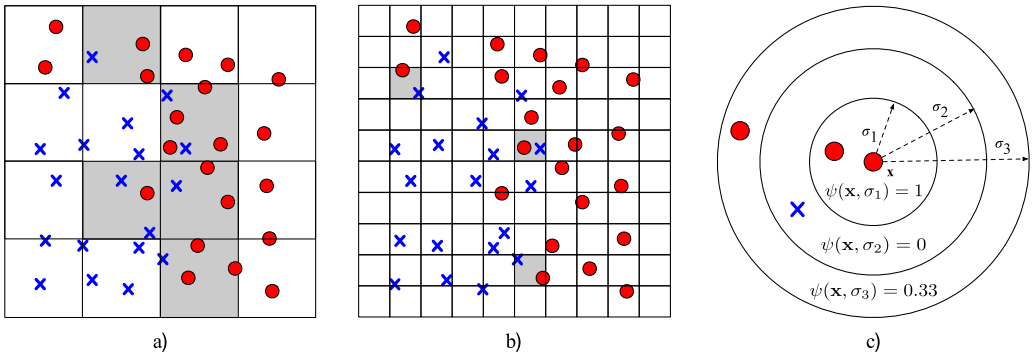

Fig. 9. Example of multiresolution overlap measures, which aggregate global and local information on the domains. In (a) and (b), a strategy of recursive feature space partitioning is used to analyse the domains at increasingly lower resolutions. At each resolution, problematic regions (grey cells) are individually analysed. In (c), example **x** exhibits distinct complexity values depending on the resolution of its neighbourhood (defined using hyperspheres with different radii). The final characterisation of domains consists of averaging the individual results obtained at several resolutions.

3.3.1. Properties of the proposed taxonomy

Beyond mapping the relationship between complexity measures and their associated class overlap representations, the proposed taxonomy evidences certain properties of the measures and illustrates other existing relationships between the categories that constitute the taxonomy. In particular, three main characteristics may be highlighted:

1. Measures belonging to different decomposition or identification categories may be associated to the same class overlap representations: As shown in Fig. 3, there are situations where measures based on distinct decomposition and/or identification strategies aim to provide similar insights. An example is the case of Purity and Neighbourhood Separability measures, C1 and C2, and MRCA, which are comprised in the "Multiresolution Overlap" group (since their insights are derived from the same underlying principle), despite the fact that their identification of problematic regions is performed differently (through "Feature Space Partitioning", "Neighbourhood", and "Hypersphere Coverage", respectively). The same rationale applies to other examples depicted in Fig. 3.

This evidences that the strategy through which overlapped regions are decomposed and identified, may not correspond directly to the knowledge they incorporate. In other words, this illustrates that although the analysis of the process of decomposition and identification of problematic regions is essential to the characterisation of class overlap, investigating its quantification and the insights provided by each complexity measure – through a careful analysis of their design and purpose – is fundamental to fully understand the problem. To some extent, existing research has often grouped complexity measures according to the process inherent to the identification of certain properties (e.g., feature-based, neighbourhood-based) [40,41], rather than the insight they produce on the data domain. In this regard, one of the advantages of the presented taxonomy is that the decomposition and identification processes of each measure can be dissociated from the perception obtained from data, i.e., measures are grouped based on the knowledge they provide on the domain, rather than on their underlying processes. Nevertheless, such information is not lost, since it remains established in the upper-levels of the tree structure that compose the taxonomy.

2. Measures may incorporate two or more decomposition or identification methods: Although the established groups are subsets of complexity measures with shared similarities, their boundaries are not strictly delimited. Accordingly, some measures may comprise two or more decomposition or identification methods. To some extent, they may be considered "hybrid" measures, which is the case of N1 and DBC. N1 is based on graph decomposition although it also incorporates neighbourhood information to identify connected vertices with disagreeing class memberships. In turn, DBC first divides the domain into hyperspheres, and then builds an MST considering their centres and analyses the neighbourhood of the MST vertices. Both their insights are however more related to boundary complexity and the internal structure of classes (structural overlap) rather than to local data characteristics (neighbourhood analysis) and are therefore included in the Structural Overlap group.

3. Measures that complement certain representations of class overlap: Some groups of measures are also intrinsically related to (or complemented by) others, as previously discussed. This is the case of Instance Overlap measures, that cannot be dissociated from the concept of "Instance Hardness", and the case of Structural Overlap measures, which encompass the characterisation of the "Density of Manifolds". We have chosen to highlight these two subgroups in the taxonomy since, notwithstanding their representations, they are often crucial to devise optimal solutions for certain domains. When analysing the current panorama on class imbalance and overlap problems (Section 4), we will see how instance hardness information is useful for preprocessing approaches, and often embedded in the internal operations of some resampling algorithms for imbalanced learning. In turn, instance overlap measures provide a better insight of the overall difficulty of the domain for classification. Similarly, some class overlap-based methods, more than analysing certain global properties of the domains (e.g., structural properties), may further incorporate density information for improved results.

3.3.2. Sensitivity of complexity measures to class imbalance

Another topic of discussion is whether the identified class overlap measures are sensitive to class imbalance. In Fig. 3, class overlap measures that have been designed or adapted to be attentive to class imbalance are marked with an asterisk.

Some measures take the problem of class imbalance into account by defining the data typology only for the minority class (Borderline Examples [61]). Others were originally proposed within the scope of imbalanced domains (R_{aug} [38], ONB [41], CM [34], wCM and dwCM [17]), although only R_{aug} incorporates the imbalance ratio in the computation of class overlap (the remaining use a strategy of class decomposition, i.e., complexity measures are computed for each individual class). The same applies to recent adaptations of well-established measures (F2, F3, F4, N1, N2, N3, N4, T1), also based on class-wise computation [42].[4] The basis for the development of these adaptations is that, in imbalanced domains, the majority class tends to dominate the computation of some complexity measures, providing biased estimates of classification difficulty [34,54,69]. This is mostly observed for measures that depend on the total number of examples in data, rather than class sizes. Ongoing research therefore considers the

[4] In this regard, F1 was also studied in [42,54], although, since it relates two means and variances, it was not possible to adapt it in order to obtain individual information by class. The same is expected for F1v.

decomposition of complexity measures into their minority and majority counterparts, and has shown promising results for binary-classification tasks [42,54] (this will be further discussed in Section 4).

Other than the highlighted measures, the remaining are yet to be investigated in the context of class imbalance.[5] A final remark should be given to MRCA [37], that although it was not especially devised or thoroughly investigated in imbalanced domains, it considers an *imbalance estimation function*, which attends to the distribution of examples comprised within the hyperspheres, at each step, before obtaining a complexity profile of a given example.

3.3.3. Implications for future research

Let us now delve into the implications associated with the inception of our proposed taxonomy for future research in the field.

In alternative to discussing general measures of classification complexity, our taxonomy focuses specifically on class overlap. Among well-known data issues, this is the most harmful for imbalanced learning tasks [5,9] and the one which generates most debate regarding its definition, measurement, and understanding [18]. In this regard, the proposed taxonomy clarifies the concepts associated with the definition, identification, quantification and characterisation of class overlap, and illustrates its distinct representations, as well as the sources of complexity to which they are associated.

Additionally, rather than aggregating complexity measures solely according to the category of data descriptors (e.g., separability, topology, sparseness, decision boundary) or their object of study (e.g., feature-based, neighbourhood-based, network-based), the proposed taxonomy focuses on associating class overlap measures to the insight they provide regarding the domain. In other words, each measure is associated to the class overlap representation it is able to perceive. Consequently, several practical implications for future research may be drawn:

- The proposed taxonomy advocates for the establishment of standard measures of the overlap degree, contrarily to what is still currently portrayed in related research, where class overlap is measured in rather distinct ways.[6] To this regard, the proposed taxonomy evidences which measures are better suited to capture specific types of class overlap, should the researchers be interested in a particular facet of the problem;
- Notwithstanding the effort to associate each measure with the class overlap representation it captures, the proposed taxonomy reflects simultaneously the three basal components of class overlap characterisation (decomposition, identification and quantification/insight). As such, it allows that different groupings are established depending on the intended level of the analysis;
- Acknowledging class overlap as a heterogeneous concept, our taxonomy further advocates for the need of a complete characterisation of class overlap, through the combination or simultaneous analysis of distinct representations of the problem. In this regard, the properties and relationships between measures identified in the taxonomy may serve as a stepping stone for the development of more perceptive, flexible and robust sets of complexity measures;
- Beyond well-established measures, this taxonomy includes more recent (although lesser-known) measures, often encompassed in uncharacterised groups (e.g., “Other Measures” [40]). The new taxonomy actively characterises their properties, relationships and insights, which contributes to a broader and deeper knowledge on the topic;
- The taxonomy also identifies class overlap measures that have been developed in the scope of imbalanced domains, or for which adaptations to imbalanced data have been explored in the literature. Accordingly, it illustrates to which extent the joint-effect of both issues has been discussed in the scope of classification complexity, and highlights opportunities for novel contributions in the field.

To summarise, the proposed taxonomy systematises the current state of knowledge of the problem of class overlap in what concerns its definition, identification, quantification and characterisation. Furthermore, it highlights core properties of the measures and provides an overview of the relationships between them. Finally, it evidences that future research should keep moving towards the development of measures with broader points of view, i.e., that are able to combine different representations of class overlap and consider other factors, namely class imbalance.

Along the next sections, we offer a multi-view panorama of the state-of-the-art solutions for class imbalance and overlap across several branches of machine learning. The main goal is to analyse the current body of knowledge in different but related areas of research, identify their limitations and suggest possible future directions. Whenever possible, insightful class overlap measures are identified and discussed within each area, based on related research on the respective topics.

4. Class imbalance and overlap: A multi-view panorama

In this section, we summarise how state-of-the-art research tries to handle class imbalance and overlap jointly across different fields, taking into consideration the ideas discussed throughout Sections 2 and 3. To provide the reader with a global understanding of the current state of knowledge, Fig. 10 illustrates the main topics discussed throughout this section. Four main areas (and respective sub-areas) of research are identified and will be presented following the schema of Fig. 10, moving from the top-left corner to the lower-right corner: Data Analysis, Data Preprocessing, Algorithm Design, and Meta-learning. Herein, we focus mostly on the topics that are currently being explored more thoroughly within each field, summarising their most significant insights. In light of the class overlap representations and taxonomy previously presented, we provide a discussion on insightful complexity measures for each topic, whenever possible: naturally, some topics will be more deeply supported by the use of complexity measures than others. Finally, although we provide a general view on all topics in Fig. 10, those that are investigated less often are marked as open challenges and will be further discussed in Section 5, where we present promising lines for future research, explaining how the considerations of Sections 2 and 3 could lead to improved solutions.

4.1. Data Analysis

One of the most prominent use of complexity measures is their application to establish the baseline classification difficulty of a given dataset. Insightful complexity measures produce estimates that are aligned with the performance of classifiers, i.e., by determining complexity measures over different datasets, we may infer which will yield better classification results. Overall, class overlap measures have proven to be good indicators of classification difficulty, although imbalanced domains require a more thoughtful characterisation given their observed bias towards the majority class [42]. Data analysis is perhaps the most frequently studied topic on the problem of class imbalance and

[5] Note, however, that according to previous results, a biased behaviour is expected for complexity measures that provide average values over the total number of observations. Nevertheless, they are simple to adapt through class decomposition.

[6] For instance, some works refer to specific measures (F1 [70], N1 [71], or data typology [72]), while others refer to a generic Overlapping Ratio [13,49,50], which is based on different variations of instance overlap measures. Besides not using a standard measurement of class overlap (and hence preventing a fair comparison between approaches), related work is in fact focusing on distinct facets of class overlap, by resorting to measures that capture different dimensions of the problem.

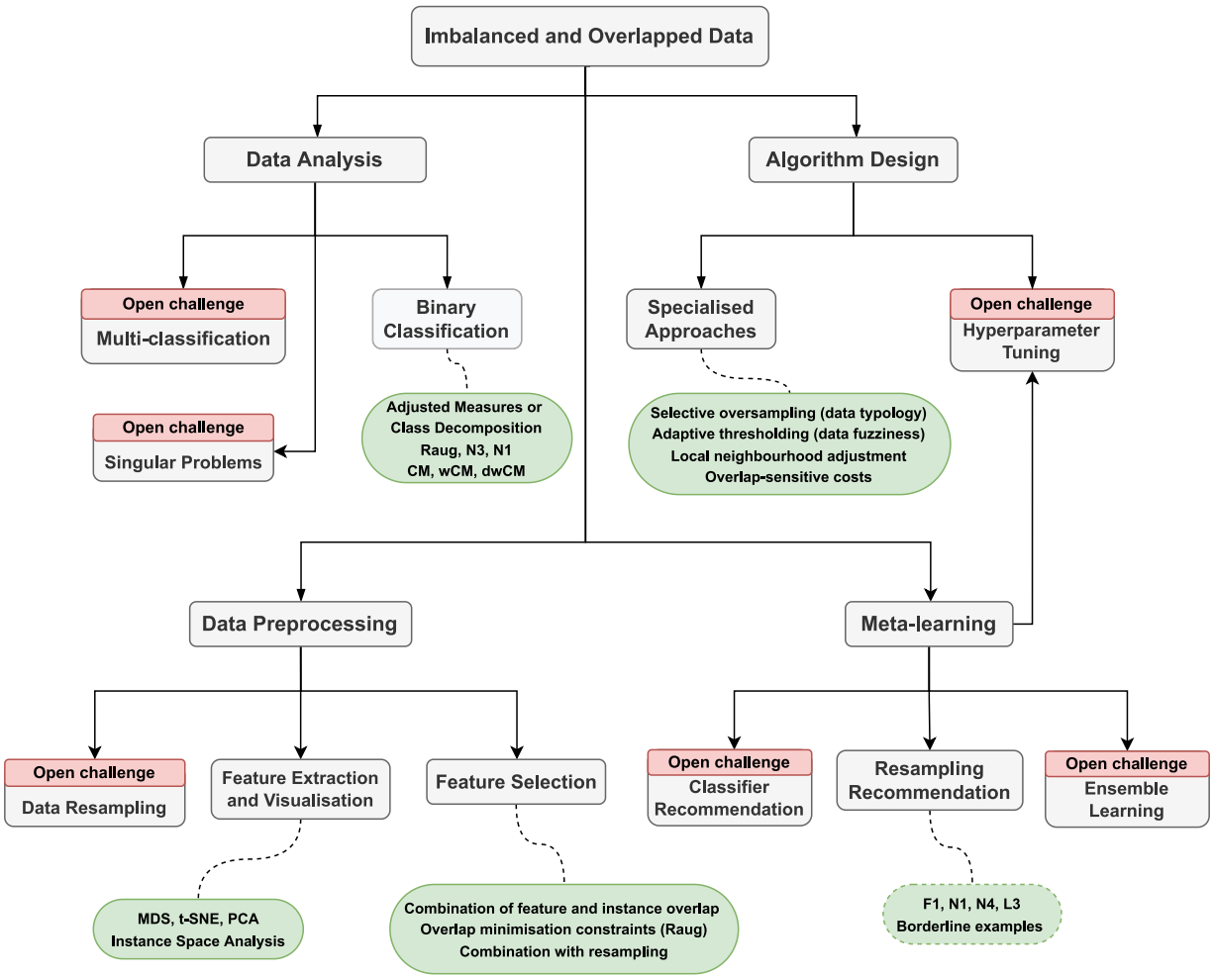

Fig. 10. Overview of current research in imbalanced and overlapped domains. Underinvestigated topics are identified as open challenges, whereas for the remaining, the major insights for research are summarised. Whenever relevant, insightful class overlap complexity measures are also highlighted, based on the findings of related research on the topic.

overlap, where different lines of thought are currently under investigation, depending on the classification paradigm. For binary-classification problems, the current established approach relies on the decomposition of complexity measures by class, whereas multi-classification and singular problems present additional challenges for research. In what follows, we will detail the state-of-the-art recommendations when handling these scenarios.

4.1.1. Binary classification

In binary imbalanced domains, the majority class tends to dominate the computation of some complexity measures [34,69]. The focus is therefore shifting towards the proposal of adapted measures that incorporate class imbalance or the evaluation of the individual class complexities, i.e., decomposing complexity measures into their minority and majority counterparts [34,38,41,42,54].

Related research has demonstrated how several of the complexity measures by Ho and Basu are insensitive to class imbalance in overlapped domains and propose new complexity measures that correlate better to the classification performance of the minority class (e.g., R_{aug}) [38]. Another line of research is the adaptation of the original measures by Ho and Basu [42,54], where complexity estimates are provided for the majority and minority class individually, rather than taking a single measure for the entire domain.

In particular, instance overlap measures (please refer to Section 3.2.2) have demonstrated an exceptional good alignment with classification difficulty, with adaptations of N3, CM, wCM and dwCM for the minority class obtaining the highest correlations with performance results [17,34,42]. Instance hardness measures have also proven to be good estimators of classification complexity [2,8,61]. As they look for hard examples to classify, it is intuitive that they are the very aligned with classification performance. In particular, measures that relate to class overlap (kDN, borderline and rare/outlier points) have been identified as major contributors to classification difficulty. Note how the most useful complexity indicators are highly correlated: it becomes clear that analysing the local properties of the domains is a suitable approach to determine classification difficulty in the case of imbalanced binary-classification domains.

4.1.2. Multi-classification

Contrary to binary-classification problems, a decomposition by class may not suffice to accurately estimate the difficulty of the classification tasks in multi-class domains: previous research has shown some inconsistencies between the complexity obtained for a given class and the performance achieved on that class [58]. Nevertheless, the co-decomposition of complexity measures considering the combination of existing classes may be used to characterise multi-class domains more deeply. In particular for class overlap, this may be helpful to establish which classes have broad overlapping areas with the remaining or which classes are responsible for the most problematic areas.

Another advantage of co-decomposition is the ability to integrate the individual properties of classes in the computation of a final measure. For instance, R_{aug} (Table 1) could be used to measure the overlap of every two classes, where the imbalance between those classes will

also be captured. Alternatively, previous class-wise adaptations of complexity measures may be further examined in multi-class imbalance domains, i.e., determining the complexity between every two classes.

The major question here is how to determine an overall measure for the entire domain, which constitutes an open issue for research. Most frequently, strategies to compute complexity measures over multi-class datasets rely on One-Versus-One (OVO) or One-Versus-All (OVA) approaches. OVO considers all possible combinations for every two classes in the domain, i.e., $\binom{C}{2}$ binary sub-problems (C representing the total number of classes in the domain). In turn, OVA tests every class against the remaining, composing C binary sub-problems. In both cases, a final measure may be defined as the average across all sub-problems. This is in fact the default behaviour of existing software for complexity measures: `DCoL`,[7] uses OVA whereas `ECoL`[8] `ImbCoL`[9] and `pymfe`[10] use OVO. However, this type of decomposition somewhat perverts the decision boundaries of the original domain, since the individual properties and relations between classes are disregarded.

Naturally, more thoughtful measures such as R_{aug} or the adaptations of complexity measures allow to incorporate more information into the final measure, namely the imbalance between classes, thus avoiding treating all pairs of classes equally. Similarly, it is possible to define several approaches for the aggregation of individual values (rather than the average). One possibility is to weight the contribution of each class to the overall overlap according to the representation of the class concept in the domain. Other possible aggregations have recently been derived [16]. Despite that, new approaches need to be investigated, especially taking into account the mutual relationships between classes. Possible directions are to consider cluster-based solutions [49] or incorporating the similarity between classes while computing data typology [73]. We acknowledge this topic as one of the major issues for future research and discuss some approaches for multi-classification domains in Section 5.1.

4.1.3. Singular problems

The great majority of studies in the field of imbalanced learning is focused on standard supervised learning tasks (often classification tasks, either binary or multi-class). With respect to non-standard supervised learning problems, i.e., singular problems, little research has been developed and therefore their study constitutes another challenge for future research. Singular problems comprehend a set of variations of non-classical supervised learning problems, where the traditional structure (e.g., one-vector input and one-dimensional output) does not apply. This is the case of multi-label, multi-instance, and multi-view problems, to name a few [74]. Complexity measures have the potential to be as useful for singular problems as they have been for standard classification problems. Nevertheless, similarly to what has been recently uncovered for imbalanced domains (i.e., that several complexity measures are biased towards the most represented concepts), they require further adaptations to properly handle problems with a different composition. A recent review on singular problems may be found in [74]. A discussion of recent research related to class imbalance in the singular problems framework is presented in Fernández et al. [43]. With respect to imbalanced and overlapped domains, Pascual-Triana et al. [41] describe some strategies to adapt ONB measure (representative of structural overlap) to several types of singular problems. Possible future directions within the scope of singular problems are discussed in Section 5.2.

[7] https://github.com/nmacia/dcol
[8] https://github.com/lpfgarcia/ECoL
[9] https://github.com/victorhb/ImbCoL
[10] https://github.com/ealcobaca/pymfe

4.2. Data Preprocessing

Data Preprocessing encompasses a series of operations that may be applied before the data is passed to the learning stage, where the classification models are built. In the context of imbalance and overlapped domains, common preprocessing tasks include:

- Data Resampling: To compensate for class imbalance by removing majority examples and/or synthesising new minority examples, and to identify and clean overlapped regions or examples;
- Dimensionality Reduction: To alleviate the dimensionality ratio problem (i.e., the *curse of dimensionality* [75]), by characterising the data domain through a reduced representation, rather than the entire input data. This process is commonly performed using feature selection (selecting a subset of the original features by discarding redundant and/or overlapped features), or using feature extraction (replacing the original features with new transformed/extracted features that retain the relevant information in data);

4.2.1. Data Resampling

In this section, we focus on the current trends on handling imbalanced and overlapped domains. To that regard, Fig. 11 summarises the most popular approaches in the field, along with the class overlap representations (introduced in Section 3.2) they are associated to. The reader may find additional information on class overlap-based approaches in [19]. Among class overlap-based approaches, data resampling approaches (undersampling, cleaning and oversampling) are the most frequently explored when handling class imbalance and overlap simultaneously. Nevertheless, when relevant, we also provide some comments on the remaining approaches.

In light of the class overlap representations previously discussed, it is possible identify some trends regarding the development of approaches sensitive to class imbalance and overlap. Undersampling approaches are more prone to consider structural information, via clustering and graph-based approaches [63,64,76,77]. They focus on defining the regions of interest (core concepts) of the data domains and discard redundant or overlapped examples found within those regions. In turn, cleaning and oversampling approaches mostly prioritise local information, often via kDN rules [78]. In cleaning approaches, the value of k determines the depth of the cleaning procedure (either addressing borderline regions or the entire domain). In this regard, multi-resolution information (fine-grain search) information has been explored to recursively remove harmful examples from data [79]. Oversampling is increasingly moving towards parametrised approaches that adapt the generation of new examples to the characteristics of each dataset [80–83]. There is also some concern with the generation of examples that are both informative and diverse [84,85]. This allows the generation process to cover more regions of the data space and alleviate the structural complexity of datasets to some extent. Oversampling approaches therefore seem more flexible, but may require a large number of user-defined hyperparameters, for which there is not yet an established relationship with complexity measures. This constitutes yet another open challenge for hyperparameter tuning (more details in Section 5.4). Finally, it is not uncommon for approaches to share some paradigms or consider several sources of information (e.g., local, structural, density, fuzzy logic, cost-sensitive). This goes towards the idea that class overlap has different vortices of complexity and addressing them altogether could potentially improve results. Additionally, there are considerably fewer approaches developed within the scope of ensembles, evolutionary, region splitting and hybrid approaches. This may be due to the lack of current knowledge on the joint-effect of class imbalance and overlap on different learning paradigms [15], ensemble learning, and hyperparameter tuning.

Despite the fact that some research has been focused on handling domains affected simultaneously by class imbalance and overlap in

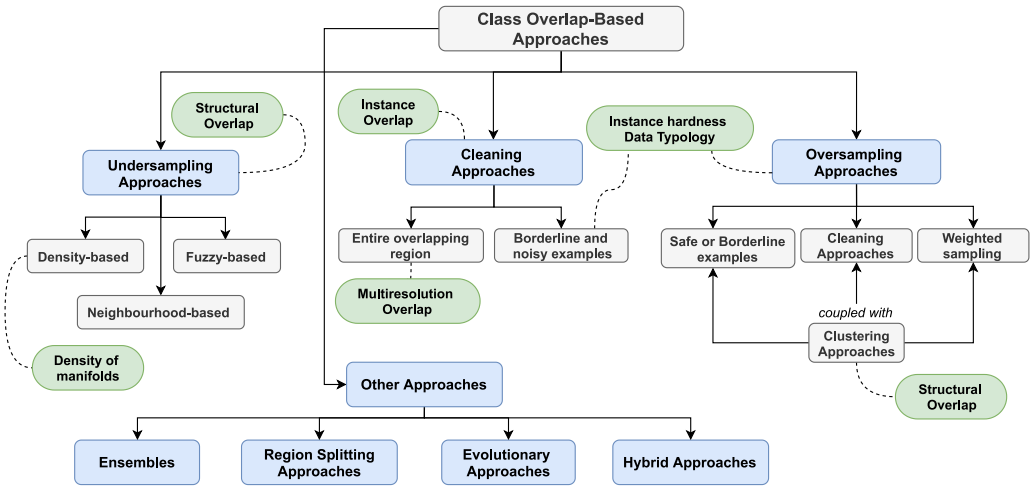

Fig. 11. Common approaches to address imbalanced and overlapped domains. The schema associates each group of approaches to the class overlap representation it is most attentive to.

the last couple of years, there is currently not enough knowledge to support the application of one approach (or category of approaches) over the others. On the one hand, despite the extraordinary flexibility of oversampling methods, the generation of synthetic examples becomes a more complicated task in overlapped domains due to the risk of further exacerbating class overlap, i.e., generating examples in problematic regions [10]. This has been somewhat attenuated by the development of more polished approaches [81–85], but at the cost of increasing computational complexity and interpretability (too many user-defined hyperparameters to tune). On the other hand, the apparent superiority of oversampling techniques due to their ability to consider the inner structure of data [86] may not hold for imbalanced and overlapped domains. Indeed, most recent undersampling and cleaning approaches also consider information regarding the structural and local complexity of the domains and have proven to surpass well-established oversampling algorithms [63,64,77]. Additionally, there are obvious advantages to using other types of approaches, such as the incorporation of data complexity and classification performance in multi-objective evolutionary approaches [50,71], or the combination of multiple reasoning paradigms when using ensembles [31,49].

Beyond a theoretical point of view, there are further empirical limitations preventing recommendations on the best approaches to handle imbalanced and overlapped domains from being devised. These relate to experimental design of related work, the lack of a standard definition, characterisation, and measurement of class overlap, as discussed along Section 2, and the lack of dataset benchmarking and open software. We will discuss these limitations and directions for further research in Sections 5.3, 6.1, and 6.2.

4.2.2. Feature selection

Feature Selection is an important preprocessing step when handling high-dimensional data in every standard classification domain, given that a large number of features can be problematic for some classifiers [87]. In imbalanced and overlapped domains, it becomes a more strenuous task since it is more difficult to discriminate certain concepts in data and consequently determine the features that increase class separability.

Past work has already discerned on the challenges of feature selection in imbalanced domains [43], whereas the use of complexity measures for the recommendation of feature selection methods has become a hot topic in the last couple of years. Okimoto et al. [88] show the suitability of using data complexity measures for univariate feature selection, where F1, F3 and N1 were successful in selecting the most relevant features. F1, associated with class separability was the most effective. In a later work, F1 is coupled with N2 to produce a univariate–multivariate feature selection approach [89], combining both feature-based and neighbourhood-based information. Parmezan et al. [87] proposed a new framework for the recommendation of feature selection algorithms based on meta-learning, considering both the characteristics of the feature selection methods and the intrinsic characteristics of the datasets. Information theoretic and complexity meta-features have shown promising results in the characterisation of datasets [44]. In particular, the ratio signal/noise, dispersion of the data set and average mutual information between classes and attributes were frequently selected as decision nodes in the meta-models. Similarly, F2 was also present in the all the constructed meta-models. Seijo-Pardo et al. [90] use a combination of feature overlap measures (F1, F2, F3) to guide the definition of thresholds regarding a suitable number of features to keep by feature selection methods. Dong and Khosla [91] show that the performance of feature selection methods is correlated with N3.

A few emergent approaches have attempted to handle class imbalance and overlap in synergy. Fernández et al. propose a multi-objective evolutionary algorithm to handle class imbalance and overlap [92]. Both feature and instance selection are considered while evolving solutions, to simultaneously compensate for the class distributions, remove complicated examples, and remove features with high overlap degrees. Lin et al. [93] propose a feature selection algorithm based on feature overlapping and group overlapping (FS-FOGO). Feature overlapping is computed by the ratio of the overlapping region on the effective range of each class (similarly to F3), while group overlapping is determined by the number of examples that fall onto overlap regions between classes (using R-value [58]). In such a way, *group overlapping* is related to the instance overlap category defined in Section 3.2, and FS-FOGO combines it with feature overlap to better decide on the discriminative power of features. Fu et al. [16] propose two feature selection methods to define a subset of features under SVM and Logistic Regression classifiers: MOSNS (Minimising Overlapping Selection under No-Sampling) and MOSS (Minimising Overlapping Selection under SMOTE). Both methods are built via sparse regularisation with the main objective to minimise the overlap degree between the majority and the minority classes (defined using R_{aug}, therefore incorporating instance overlap information). However, MOSS first applies SMOTE to rebalance the training data. MOSS outperform all other approaches (MOSNS, ACC and ROC-based feature selection) regarding classification performance, whereas MOSNS produces the lowest number of retained features while providing better or comparable results to ACC and ROC-based methods in most datasets. Recently, MOSS has also shown to improve the performance of imbalanced approaches in multi-class domains [94]. Based on the same strategy of considering sparse

feature selection to minimise class overlap (i.e., instance overlap, via R_{aug}), Fatima et al. [95] refer to RONS (Reduce Overlapping with No-sampling), ROS (Reduce Overlapping with SMOTE), and ROA (Reduce Overlapping with ADASYN). RONS and ROS are the same as MOSNS and MOSS, respectively, while ROA follows the sample principle as MOSS although using ADASYN instead. Considering ADASYN instead of SMOTE seems favourable, since ADASYN focuses on more complicated minority examples, whereas SMOTE considers all minority examples equally.

4.2.3. Feature extraction and visualisation

Rather than selecting a subset of features, feature extraction methods perform certain transformations on the original set of features in order to produce a reduced set of artificial features. These new features are somewhat a combination or mixture of the original features that aims to retain most of the information comprised in the original feature space. In imbalanced and overlapped domains, a common application of feature extraction is data visualisation. Graphic inspection is often applied to get a feel of the structure of data, overlapping between classes and overall data complexity. To that end, datasets are often transformed using feature extraction techniques to allow data visualisation in two or three dimensions.

Anwar et al. [34] used Multidimensional Scaling (MDS) to represent each data example in two dimensions in order to visually assess data complexity. The visualisation is used in conjunction with the proposed CM metric (Table 1) to analyse the degree of overlap between classes. Whereas the majority class is shown in some colour, each minority class example is identified by the number of same class neighbours in its 3-neighbourhood. Napierala et al. [61] used MDS and t-Distributed Stochastic Neighbour Embedding (t-SNE) to assess the dominant typology of datasets (safe, borderline, rare/outlier datasets) and identify class overlap. Despite certain differences in the projections of both methods, the observations regarding the complexity of the studied domains are similar.

Recent research is also exploring feature extraction and visualisation strategies to characterise the footprint of algorithms. This is a methodology known as *Instance Space Analysis* and may be applied to a collection of datasets or to individual observations within a dataset. The rationale of the analysis is similar. Essentially, it involves summarising each dataset or each instance within a dataset as a n-dimensional feature vector representing its complexity. Regarding the taxonomy presented in Section 3, dataset complexity may be captured by the class overlap representations proposed, where the complexity of singular observations are most often associated with the instance hardness measures [8]. Then, using a feature extraction technique, e.g., Principal Component Analysis (PCA), a two or three dimensional embedding (an *instance space*) that can be visually investigated. The classification performance associated to each dataset or instance can be superimposed in the visualisation to identify regions of good or poor behaviour of classifiers, and identify pockets of hard and easy datasets or instances. Smith-Miles et al. [96,97] used PCA to project dataset instances onto a 2-dimensional space and analyse algorithm performance. Muñoz et al. [98,99] propose a new dimension reduction methodology that improves the interpretability of the visualisations. The new projection approach is optimised so that the created instance space represents as much as possible a linear trend between data complexity and classification performance.

4.3. Algorithm Design

The idea behind algorithm design is to adjust a given approach, i.e., the parameters of a classifier or preprocessing method, to the characteristics of data. In the context of imbalanced and overlapped datasets, a common strategy is to incorporate information regarding both these problems in the development of approaches. Such information might appear in the form of an heuristic based on complexity measures and/or other observed characteristics of datasets, leading to the development of specialised approaches. Alternatively, it can also be based on the tuning of hyperparameters. In this case, the main objective is to maximise the classification performance by choosing optimal hyperparameters for classifying or preprocessing each dataset.

Whereas some strategies for specialised approaches have been applied in the literature, hyperparameter tuning remains an understudied topic in what concerns the design of approaches sensitive to the peculiarities of data suffering simultaneously from class imbalance and overlap.[11] In what follows, we discuss some existing approaches in this regard.

4.3.1. Specialised approaches

Depending on the category of class overlap based-approaches (please refer to Section 4.2.1), different strategies may arise for the development of specialised approaches. Recent approaches are based on defining heuristics for undersampling or cleaning (adaptive thresholding or local neighbour adjustment), analysing local information for selective oversampling (via data typology) and incorporating costs associated with data complexity directly into the learning systems.

Pattaramon and Elyan [64,79] propose two heuristics for cleaning overlapped majority class examples. With AdaOBU [64], they introduce an automatic elimination threshold adaptable to the degree of class overlap. The threshold is proportional to the fuzziness of the dataset and consequently to the existing class overlap. In [79], authors discuss another heuristic to determine a reasonable value of k for neighbourhood-based cleaning methods that promotes the discovery of overlapped majority examples. The heuristic considers information regarding both the number of examples in data and the imbalance ratio. A similar approach is taken in [101], where k is defined by the imbalance ratio of the dataset.

Data typology has also been considered in the design of specialised approaches, where selective oversampling has proven to improve classification results. Skryjomski et al. [102] show how SMOTE can be empowered by incorporating information regarding the typology of minority class examples. Similarly, Sáez et al. [103] guide the oversampling procedure based on the data typology of examples in multiclass datasets. The best oversampling configurations often involved the oversampling of only borderline and outlier examples, with a higher frequency of the preprocessing of borderline examples.

Another strategy is to integrate the information regarding data complexity directly on the learning stage of classifiers. Lango et al. [72] suggest to consider the information produced by ImWeights regarding the number of clusters and associated difficulty (incorporating both structural and local information). Lee et al. [13] introduce the concept of overlap-sensitive costs, which combines both the imbalance ratio and the degree of overlap of training observations (based on kDN).

4.3.2. Hyperparameter tuning

Hyperparameter tuning allows to determine specific model parameters tailored to the characteristics of each dataset in order to obtain optimal performance. Thus, more than embedding "rule of thumb", theoretical settings into the approaches, it is possible to empirically fine-tune parameter values for individual datasets, improving classification results.

[11] Note that hyperparameter tuning, *per se*, constitutes a topic of interest across several fields beyond traditional Supervised Learning, such as Deep Learning, and Meta-learning [100]. Accordingly, some intersections between terms, trends, and solutions are likely to arise. Notwithstanding, in this paper, we detach from that intersection and overall considerations on hyperparameter tuning regarding the Deep Learning and Meta-leaning fields specifically. In alternative, we focus particularly on hyperparameter tuning with respect to imbalanced and overlapped domains, highlighting existing limitations which are yet to be addressed by all communities.

With respect to imbalanced and overlapped domains, the tuning process is most often performed directly by analysing the effect of hyperparameters on classification performance [49,62,70,80,84,104]. That involves testing a range of hyperparameters (or combinations of hyperparameters) over a benchmark of datasets and choosing the one that performs the best overall.

Some studies further discuss the effect of hyperparameters of the proposed approach and suggest appropriate values that provide overall good results. This is especially the case of approaches that require several user-defined hyperparameters (e.g., A-SUWO, NI-MWMOTE, IA-SUWO) [81–83]. Still, the discussion is given as a high-level view of the approach, rather than providing recommendations based on data characteristics. An exception can be highlighted for Douzas et al. [85], where some hyperparameter recommendations for G-SMOTE are given based on the imbalance ratio, and the ratio of the number of samples to the number of features of the datasets. Another important exception are evolutionary-based approaches that, by recurring to multi-objective algorithms, are able to consider both the classification performance and data characteristics in the refinement of the approach [71,105].

Nevertheless, there are still several approaches where hyperparameters are defined according to the default values of existing software packages or set to common values for consistency with other works in the literature that used the same approaches or datasets [76,78,106]. All in all, in what concerns imbalanced and overlapped data, hyperparameter tuning remains a neglected subject and it constitutes a challenge for further research. Accordingly, future directions will be highlighted in Section 5.4.

Finally, as previously discussed,[11] we may argue that this topic also falls onto the scope of Meta-learning and Deep Learning.

In the Meta-learning community, hyperparameters themselves may be seen as meta-data that describes the learning tasks [100], and some categories of meta-features (e.g., model-based, landmarking) further require the definition of hyperparameters as well, for which tuning is yet to be explored [44]. The idea of defining appropriate parameters depending on the data characteristics has also been subject of previous work in the field, where meta-models are designed to recommend specific configurations or hyperparameters, based on some meta-features. The reader is referred to [44,100], which constitute two comprehensive surveys on the topic. Nevertheless, existing work mostly focuses on traditional meta-features (e.g., simple, statistical, information-theoretic) rather than complexity measures, and there is not, to our knowledge, any study that focuses specifically on hyperparameter tuning for imbalanced and overlapped datasets. We will further discuss this matter in Section 4.4.

With respect to the Deep Learning field, some recent research is starting to study the behaviour of deep learning systems in imbalanced domains which are further affected by additional complexity factors, such as class overlap. The reader is referred to [107] for the first novel thoughts on the subject, although some core issues persist in deep learning systems as for their classical counterparts: class overlap remains a challenging factor even for deeper architectures, and, to this point, model parametrisation follows the same principle of experimenting with several hyperparameters to report optimal classification results.

4.4. Meta-learning

In Meta-learning (MtL), the characteristics of a dataset (named meta-features or meta-characteristics) are extracted and associated to the classification performance obtained over it. By compiling meta-information on a collection of datasets with associated performance results (thus creating a meta-dataset), it is possible to build a recommendation system that infers on the behaviour of a technique (or suggests the application of an appropriate one) based on the characteristics of a new dataset.

Traditionally, there are five categories of meta-features discussed within MtL frameworks: simple, statistical, information-theory, landmarking and model-based meta-features [108]. However, although they were not originally proposed for meta-learning, complexity measures have been used extensively in the MtL and AutoML literature [109–112]. For that reason, authors have started to refer to them as an extra category of meta-features [44] and recent research has been showing that they may prove equally or more informative than traditional meta-features [112]. In particular, class overlap measures have stood out as highly accurate indicators of classification performance [38,42]. Indeed, some class overlap measures are related to the landmarking category of meta-features. Landmarking meta-features characterise datasets based on the classification performance of simple and fast learners, such as kNN and linear discriminants, therefore highly associated with the instance overlap measures (N3) and feature overlap measures associated to class separability (F1, F1v).

In the context of imbalanced and overlapped domains, common applications of MtL systems are related to the recommendation of classifiers and preprocessing techniques or to the study of their domains of competence. Most often, related research focuses on obtaining a high level view of MtL frameworks rather than discussing informative measures [109–111]. Nevertheless, some works have attempted to connect the insights derived from complexity measures to the recommendation provided by the systems, which we discuss in what follows.

4.4.1. Classifier recommendation

In the scope of classifier recommendation, García et al. [113] use regression techniques to recommend the best classifier (ANN, DT, SVM, kNN) for a given dataset, based on their data complexity. The top most informative measures were N3 and N1, followed by N2, Density and T1. Luengo and Herrera [114] discuss an automatic extraction method to determine the domains of competence of classifiers (DT, SVM and kNN). The complexity measures regarded as most informative for the automatic extraction method were N1, N3, L1 and L2. Apart from the top informative measures, additional information may be useful depending on the nature of classifiers. That however, remains an under-investigated topic. Open avenues regarding classifier recommendation will be discussed in Section 5.5, along with ensemble learning, as they are related topics that suffer from similar limitations.

4.4.2. Recommendation of resampling approaches

Regarding data preprocessing approaches, complexity measures are often used to guide the choice of appropriate resampling techniques. Depending on the complexity of a domain, a suitable resampling strategy can be chosen by taking into account its intrinsic behaviour, i.e., how it works internally and to what extent it can alleviate certain data problems. Luengo et al. [53] analyse the usefulness of complexity measures to evaluate the behaviour of resampling approaches. F1, N4 and L3 proved informative to establish significant intervals of good and bad behaviour for different preprocessing approaches. Santos et al. [10] perform a thorough comparison of oversampling approaches for imbalanced datasets, supported by a data complexity analysis. The best oversampling techniques seemed to include structural information (cluster-based synthetisation), instance overlap information (use of cleaning procedures) and instance hardness information (adaptive weighting of examples). Costa et al. [112] use Exceptional Preferences Mining to extract interpretable rules to guide the recommendation of oversampling strategies for imbalanced datasets. Similarly to the previous work, class overlap measures were the most informative, namely measures related to structural and instance overlap (N1, N4) and instance hardness (proportion of borderline examples). Zhang et al. [111] propose an instance-based learning recommendation algorithm to determine the most suitable strategy to handle imbalanced datasets. They use complexity, landmarking measures, model-based measures and structural meta-features, although they only present a high-level view, with no specific measures discussed.

4.4.3. Ensemble learning

Although some ensemble-based techniques have been discussed within the scope of imbalanced and overlapped domains, ensemble learning is still an open avenue for research.

Current ensemble frameworks often incorporate one of two solutions. One is the coupling of ensembles with resampling and cleaning methods: recent approaches include CluAD-EdiDO [49], SPDM [31], and SPE [115]. The other is the simultaneous use of evolutionary approaches to handle the peculiarities of the domains. Most often, this involves the incorporation of some data complexity information in the objective criteria of evolutionary algorithms, in order to optimise the final performance of the ensemble. For instance, Fernandes et al. [71] discuss EVINCI, an evolutionary ensemble-based method that incorporates the N1 measure in the workflow to optimise instance selection. Fernández et al. [105] propose EFIS-MOEA, which incorporates both feature and instance selection.

The first strategy requires the understanding of which resampling/cleaning approaches are most suited to different domains, and may be supported by previous meta-learning studies on resampling approaches. The second strategy is more closely related to algorithm design, focusing on the development of specialised approaches and hyperparameter tuning to improve classification performance.

Indeed, note how both strategies do not specifically focus on ensemble learning from a meta-learning perspective, i.e., using complexity measures to define an appropriate set of base classifiers for the ensemble framework. That requires the choice of a pool of adequate classifiers to form the ensemble, which comprises both the analysis of how classifiers with different learning biases respond to the joint-effect of class imbalance and overlap, and the assessment of their combination (creating ensembles) for optimal solutions. However, as previously discussed, the link between data characteristics (i.e., complexity measures) and classifier recommendation is not yet well-established, and consequently, ensemble learning, to this extent, also remains an open challenge for research, and will be discussed in Section 5.5.

5. Open challenges and future directions for research

In what follows, we revisit the topics identified as open challenges in the previous section (Section 4) and elaborate on possible future research directions, based on the considerations of the first part of the paper (Sections 2 and 3). Such discussion constitutes the main contribution of this section.

5.1. Multi-class problems

As discussed in Section 4.1.2, the standard approach for multi-class problems consists of formulating several binary sub-problems, using OVA or OVO decomposition. On the one hand, these strategies allow the application of binary classifiers without additional modifications. Also, and especially when handling data overlap, they may simplify the original domain by focusing on sub-problems individually, thus easing the separation between classes [116]. On the other hand, this simplification is achieved at the cost of distorting the inner structure of individual classes (and original decision boundaries) and neglecting mutual relations between classes. For instance, a given class can either be considered the minority or majority class, depending on the size of the class it is being compared to. Some classes can also be more closely related (more similar) than others. With respect to class overlap, there can be a class or a subset of classes that is mainly responsible for overlapping regions, whereas other classes may have clear decision boundaries among each other. Classes may also have distinct overlapping regions with respect to each other. Regarding data typology, examples will be categorised in different types, depending on the classes considered to define their neighbourhood.

By manipulating the data internally, via OVA or OVO, the information on the intrinsic characteristics of each class is lost, which may lead to the application of methods that are not appropriate for the domain as a whole, i.e., they may hurt one class while trying to improve the representation of another. OVA can additionally introduce artificial class imbalance [49,116] whereas OVO suffers from the non-competence problem [117], i.e., when classifying new data, the predictions of all constructed OVO classifier are considered, even those of classifiers that have not been trained with examples belonging to the real class of that data. The following directions could be analysed to fully understand and explore multi-class domains:

- An interesting future direction is the exploration of cluster-based techniques. The domain is divided into several regions, where data complexity can then be assessed. For instance, clusters containing examples of only one class will not contribute to class overlap. In turn, clusters containing examples of multiple classes will be evaluated maintaining the original relationship between classes. A starter point for the investigation of this line of research is [49], where multi-class imbalanced and overlapped datasets are first clustered, before any cleaning and oversampling procedures.
- Another alternative to take into account the relationships between classes is to incorporate additional information on the data typology of different classes. Rather than considering each class in isolation and producing its typology (OVA approach) [103], recent research suggests to incorporate a similarity factor when determining the safety level of each example in data [73]. A major drawback in [73] is that it considers that similarity should be provided by the user (via domain knowledge or consulting a domain expert). As this is most often not available, a possibility to overcome this issue could be to estimate a similarity coefficient via similarity/distance functions. Another heuristic based on the imbalance ratio between class concepts has also been recently proposed [118]. It suggests that concepts with lower class imbalance are more similar to each other. We argue that associating class similarity to the imbalance ratio between classes might be too simplistic and suggest that the overlap degree between classes could be used in alternative, to produce a more realistic measure of class similarity.

5.2. Singular problems

As pointed out in Section 4.1.3, current real problems and applications are showing a more complex structure with respect to the classical supervised and unsupervised tasks: essentially, in what concerns their input and output variables [74]. In what follows, we discuss how problematic regions and/or instances may be identified in non-standard scenarios such as multi-label, multi-instance, and multi-view problems.

- To address multi-label or multi-domain learning, two different approaches are likely to be applied [119,120]. On the one hand, to transform the dataset into standard “single-output” problems. On the other hand, to adapt or design the classifier to cope with this type of data. In both cases, the occurrence of the imbalance and overlap issues is especially relevant, as there is a significant increase in the number of labels and combinations among them. To address the former situation, binarisation and probabilistic classification algorithms may be explored to ease the discrimination among groups of labels by simplifying the original problem. With respect to binarisation techniques, similar considerations can be taken as for multi-class problems (see Section 5.1).
- In the multi-instance paradigm, input examples are represented in groups or the so-called *bags* [121]. Every instance shares the input space, but the number of elements in a bag can be different. The final objective is identifying the class of the bag by labelling all instances associated to it, i.e., the bag is “positive” if there is at least one positive instance. In this scenario, considering the bags as “instances” by aggregation mechanisms, that is, considering

a single representative element for each one of them, eases the definition of overlap to follow the standard case. Otherwise, feature vectors should be used separately, instance by instance, possibly inducing a higher degree of overlap and complexity to the problem. This oversimplification of the problem may have an influence in the quality of the model to be obtained. In addition, few research studies still consider the event of imbalance in this context [122,123]. As such, there is a need of establishing the proper preprocessing approaches to cope with both overlap and imbalance taking into account the properties of the positive and negative bags.

- Finally, multi-view problems are defined as those in which each instance has a fixed number of feature vectors that can vary in type and format [124]. As there are different "input-spaces", the degree of overlap may vary for each of them. This implies that the characterisation of a given instance must be considered under the perspective that better establishes the separation among other labels. Multi-view problems are mainly addressed via auto-encoders and feature transfer [125], so that creating non-linear combinations of the original features for a higher-level representation of the data may lead to simpler decision functions.

5.3. Data resampling

In Section 4.2.1, we have provided an extensive discussion of the limitations and opportunities for future research regarding class overlap-based methods. Besides the ones previously highlighted, the following open directions are crucial for the development of new approaches dedicated to handle imbalanced and overlapped datasets:

- For the most part, the comparison of class overlap-based methods remains limited to well-established approaches (e.g., ROS, RUS, SMOTE, Safe-Level-SMOTE, Borderline-SMOTE) which have been frequently outperformed. It is also not uncommon to find that some class overlap-based approaches are compared to their analogous distribution-based approaches. It would be crucial to compare new methods with emergent, state-of-the-art approaches, developed for the same purpose, to provide a more accurate evaluation of results.
- Despite the fact that many methods are being proposed to overcome class overlap, there is a clear lack of information on how datasets are affected by this problem (there is no quantification of class overlap). The question of whether the applied methods provide true improvements with respect to class overlap therefore remains. Most often, approaches are evaluated in terms of classification performance, which may not be sufficient to validate the approach. It is important that future research considers a deeper characterisation of domains, especially if the purpose of an approach is to alleviate some data-related issue. New studies in the field should provide a more insightful characterisation of datasets beyond the number of samples, features and imbalance ratio. It is important to guarantee that a testbed is representative of the desired data issue to sustain the improvements introduced by a proposed approach.
- A large amount of class overlap-based methods is based on handling conflicting examples (e.g., borderline, noisy examples), whose identification relies almost exclusively on instance hardness measures (kDN rules). Future research could simultaneously explore other vortices of class overlap while performing this assessment. In this regard, exploring the taxonomy presented in Section 3 is a good starting direction.
- Class overlap measures can also be used to provide specialised data preprocessing so that the representation of minority examples is increased in overlapping regions. For instance, the generation of new synthetic examples can be guided in order to optimise a given complexity measure.
- Also, class imbalance should be explored beyond the characterisation of the disproportion between classes and consequently used for the definition of the undersampling/oversampling amount necessary for preprocessing techniques. Instead, it could be considered altogether with class overlap to produce new measures of complexity and further embedded in the operations of methods. Some recent work is already searching for solutions along this line, at the level of algorithm design (Section 4.3), which we believe to be the direction with the highest potential for future developments in the following years.
- Improved weighting schemes are also worth studying to adjust the complexity profile of training examples, e.g., closer neighbours, or minority class neighbours, may have a higher impact in complexity computation. This rationale can also be applied to data preprocessing approaches to provide a specialised resampling, depending on the difficulty of a given example.

5.4. Hyperparameter tuning

As discussed in Section 4.3.2, the configuration of hyperparameters (of classifiers or resampling approaches) is most often guided by the results obtained from the classification stage. Besides time consuming, this type of approach does not take advantage of information on data complexity, which is available, often at a lower cost than running entire experiments. The following directions may be explored in order to devise more insightful ways to guide hyperparameter tuning:

- Regarding resampling approaches – undersampling, oversampling and cleaning – a possibility is to guide the tuning of hyperparameters based on complexity measures. For imbalanced and overlapped domains, the hyperparameters of resampling procedures can be adjusted in a way that they alleviate class imbalance and minimise class overlap, by assessing the effects of given hyperparameters on suitable complexity measures. This can be thought out by addressing data complexity as a whole, for instance, focusing on minimising feature, instance and structural overlap. Alternatively, it is possible to address data complexity selectively, depending on the classification paradigm to be used after the preprocessing stage, i.e., focusing only on the most complicated factors for the classier at state. As an example, since SVMs can handle rather complex structures [6], one can focus solely on addressing instance overlap, removing harmful examples.
- Regarding classifier hyperparameterisation, it is possible to achieve a reduced range of hyperparameters to test by exploring data complexity at an intermediate stage. For instance, for SVMs, more appropriate combinations of C and γ can be explored depending on the characteristics of data. An obvious advantage of considering hyperparameter tuning based on data complexity is that complexity measures are often faster and simpler to compute than performing full classification experiments. Also, choosing more insightful ranges of hyperparameters allows the algorithm to converge faster, avoiding the need to test an extended set of possible combinations. In this regard, some interesting approaches have studied meta-models to determine whether or not to tune SVMs [126], or how to define appropriate sets of default hyperparameters [127]. Both research works consider general real-world domains and rely on the study of several data characteristics (meta-features), including some complexity measures (the former exploring imbalanced datasets in more detail). Although they do not focus particularly on the joint-effect of class imbalance and overlap, they may serve as a starting point to further explore hyperparameter tuning in these domains across several learning paradigms and methods, including preprocessing approaches.

- At the level of class overlap complexity measures themselves, a large number of measures relies on finding a k-neighbourhood, where the value of k is routinely set to a pre-defined value ($k = 5$ is a common hyperparameter). The same is true for data typology and several class overlap-based methods. This strategy obviously neglects the characteristics of the domains, although estimating k for each domain may be computationally expensive. Therefore, defining more insightful heuristics for setting k is a interesting direction for future work. Regarding complexity measures, some approaches suggest incrementally increasing k until the complexity stabilises [34]. On data typology, recent work discusses the possibility of tuning k and the used distance metric based on classification results of a kNN classifier [128]. On data resampling, some recent heuristics for defining suitable k-neighbourhoods are based on the degree of class overlap or the class imbalance of datasets [64,79,101].
- Similarly, adaptive methods for finding k should also be explored, where k could be adjusted to the local minority class densities. Traditionally, smaller values of k are more successful to recognise the less represented concepts in the overlap region. In turn, larger values of k benefit the more represented concepts in that region [7]. Future research could pursue the proposal of a framework able to select an optimal k value based on the local characteristics of data. In that regard, hypersphere coverage metrics could be informative to define optimal k values. For instance, examples with lower LSC require smaller values of k for correct classification.
- Future research may also focus on the investigation and optimisation of distance functions (both for specialised approaches and complexity measures). Although previous studies have shed some light on different behaviour of complexity measures and data typology depending on the used distance function [34,40,41,128], this remains a poorly studied topic.

5.5. Classifier recommendation and ensemble learning

As discussed throughout Sections 4.4.1 and 4.4.3, although previous studies have shown that the combination of class imbalance and overlap creates a challenging scenario for classifiers independently of their learning paradigms (i.e., the nature of the learned decision boundaries) [1], there is no study that thoroughly discusses this topic, focusing specifically on establishing its effects on distinct learning biases with respect to real-world domains. Related research has established some insights regarding the behaviour of local versus global classifiers [7], symbolic and non-symbolic classifiers [6] and classifiers with different learning paradigms [15]. However, these comprise artificially generated data domains, where class overlap, class imbalance and other factors (data typology, data structure and class decomposition, local data densities, and data dimensionality) are defined *apriori*. Transposing these studies to real-world scenarios is now possible due to the increasing number of complexity measures proposed and revisited in the last few years, and it would be of major interest to the research community. This would lay the foundation for the choice of baseline approaches for imbalanced and overlapped domains, as well as guide the selection of ensemble approaches. SVM and KNN have perhaps been the most studied classifiers under varying degrees of complexity [6,7,11,13], whereas establishing the behaviour of other learning paradigms remains an open challenge.

6. Open source contributions

In this section, we highlight further directions for future research that are complementary to those identified in the previous section and may contribute to their more rapid and effective advancement. The main contribution of this section consists of the identification of benchmarks and open-source software to boost new developments in the field.

6.1. Benchmark datasets

Popular public repositories (e.g., UCI,[12] Kaggle,[13] KEEL,[14] OpenML[15]) offer a diverse collection of datasets in what concerns their extrinsic complexity (number of instances, dimensionality, missing values, number of classes), though not focusing on their intrinsic complexity (class imbalance, class overlap, small disjuncts, noisy data and other data-related issues). Therefore, they lack diversity, i.e., their are not representative of a great span of complexity problems [98,129]. Regarding specific applications or data characteristics, KEEL is perhaps the most popular repository. It provides a collection of both standard datasets as well as datasets targeted to imbalance learning, detection of noisy and borderline examples, as well as singular problems (multi-instance and multi-label datasets). Nevertheless, other data complexity factors remain overlooked. An important contribution to research would be the creation of an open repository representative of data complexity problems. This would establish a benchmark for studies regarding the domains of competence of classifiers, as well as the development of specialised approaches and AutoML pipelines. The following directions could be taken in order to develop data benchmarks targeted to complexity analysis:

- Providing a complete characterisation of datasets comprised in well-known repositories and grouping datasets according to their complexity. Varying degrees of data complexity could be determined, and in particular for class overlap, the taxonomy provided in Section 3 could be helpful to divide datasets depending on their dominant overlap representation. For instance, some datasets can be structurally intertwined (structural overlap), whereas others may include a great amount of difficult examples (identified with instance overlap measures). Combinations of these factors could also be considered.
- On this note, it is important to refer to the computational complexity associated to the computation of some complexity measures. Despite the fact that they have been used extensively in MtL applications, their widespread usage may be compromised by the fact that some are computationally expensive. In this regard, an open challenge relies on the optimisation of complexity measures. As an alternative, recent research has shown that it is possible to predict data complexity measures of a given dataset using simpler, low cost meta-features as input [130], which could also be an interesting direction to explore.
- Complementary to the characterisation of datasets, a possible strategy to guide researchers on the choice of appropriate datasets to evaluate their proposed approaches could be the creation of a meta-dataset which could then be explored via clustering analysis to define groups of datasets with similar complexity. Another interesting approach is the one taken in [98] where datasets are projected onto a 2-dimensional instance space where their complexity and diversity can be visualised.
- Enhancing existing repositories with artificial data is also a possibility. Previous work suggests enhancing data repositories with the thoughtful design of artificial datasets, via evolutionary multi-objective algorithms [129]. This approach samples a real-world dataset so that the resulting set of examples optimises a set of data complexity measures. A similar approach based on class label modification is introduced in [131]. Another strategy is presented in [98], where datasets are evolved to fall onto target regions of the complexity space. Similarly, a recent and interesting line for future development is the exploration of *data morphing*, where a

[12] https://archive.ics.uci.edu
[13] https://www.kaggle.com
[14] http://keel.es
[15] https://www.openml.org

real-world dataset can be gradually manipulated to display certain meta-characteristics [132]. In this case, it would be possible to select a high complexity dataset with respect to certain properties (e.g., both structural and instance overlap) and iteratively transform a less complex dataset to exhibit gradual variations of those properties. Although manipulating the datasets artificially, these strategies aim to enrich their data characteristics while attempting to maintain the essence of real-world domains. With respect to class overlap, Sáez et al. [116] discuss a scheme to generate overlapping regions in real-world datasets.

- In alternative, artificial datasets can be used as a benchmark to improve the behaviour of approaches with respect to a particular aspect (e.g., presence of borderline examples, class-skews). The main advantage is that artificial datasets can be tailored to the needs of the experimental setup, i.e., covering specific sources and ranges of data complexity or gradually increasing data complexity. A recent line of research in this direction is [133], where a many-objective optimisation algorithm is used for complexity-based data generation.

6.2. Software and open source implementations

- Code availability is a crucial aspect for the reproducibility of results. Long-established methods are implemented in several open-source software. Some of the most popular are KEEL Software Tool[16] [134–136], WEKA workbench[17] [137], among other R[18,19,20,21] [138–141] and Python[22,23] [142,143] packages. However, most recent research work does not frequently provide open-source implementations of novel approaches on imbalanced and overlapped data. We have identified all existing resources (data and code) regarding class overlap-based approaches in imbalanced domains, so that researchers may consider them in future experiments.[24] We further encourage future researchers to make their code and obtained results publicly available.
- Existing open-source implementations of complexity measures include the DCoL (C++)[25] [39], ECoL[26] [40] and the recent `ImbCoL`[27] [54], `SCoL`[28] [130], and `mfe`[29] [144] packages (R code). There is also `pymfe`[30] in Python. Regarding class overlap measures identified in Section 3, these packages consider the implementation of the following: F1, F1v, F2, F3, F4, N1, N2, N3, N4, T1 and LSC. `ImbCoL` provides a decomposition by class of the original measures and `SCoL` focuses on simulated complexity measures, as discussed in the previous section. In order to foster the study of a more comprehensive set of measures of class overlap, we provide an extended Python library – *Python Class Overlap Library* (`pycol`)[31] – comprising all the class overlap measures included in the previous packages, plus the remaining measures described in Section 3: F1, F1v, F2, F3, F4, IN, Purity, Neighbourhood Separability, MRCA, C1, C2, N2, NSG, ICSV, T1, DBC, ONB, $Clst$, N1, IPoints, LSC, kDN, Borderline Examples, $degOver$, SI, R-value, R_{aug}, N3, N4, D3, CM, wCM, and dwCM. We are currently conducting a large experimental study over imbalanced and overlapped datasets, focusing on distinct representations of class overlap and the ability of the identified groups of class overlap complexity measures to effectively characterise them.
- Within the scope of artificially generated data, we also recommend the use of data generator described in [6], for which we provide the documentation in English so that more researchers are able to understand and configure it. Additionally, we include our example collection of generated artificial datasets, as well as visualisation modules for data typology.[32] We welcome other researchers to contribute with their own research data in order to move towards the creation of a representative repository regarding data complexity factors, beyond imbalanced and overlapped datasets.
- With respect to *Instance Space Analysis* discussed in Section 4.2.3, exploring `MATILDA` (Melbourne Algorithm Test Instance Library with Data Analytics)[33] is an interesting direction. It allows the visualisation of the distribution, diversity and complexity of existing benchmark and real-world instances, the generation of new synthetic test instances at specific locations of the instance space, and the analysis of algorithm footprints [98]. Another recent tool is `PyHard`,[34] which allows to assess the complexity of individual examples within a dataset [145].

7. Concluding remarks

As thoroughly discussed throughout this work, real-world applications need to account for both class imbalance and overlap when devising suitable solutions for domains affected by both problems. However, whereas class imbalance is simpler to characterise and measure, referring to the disproportion of examples between classes, class overlap stands as a confounding concept, due to the multitude of representations, i.e., specific types of overlap problems, it comprises. For instance, some authors may characterise overlap as the overlap between individual feature values, associating class overlap to the discriminative power of features. Others may characterise the problem by searching for complicated examples located in borderline regions between classes, in which case class overlap refers to instance complexity. The lack of a standard and well-formulated characterisation of class overlap in real-world domains is currently preventing the research community to move towards improved approaches since, due to the lack of consensus and standardisation, the evaluation (and consequently, the comparison) of existing solutions and associated results (and insights) becomes extremely difficult.

In this work, we advocate for a unified view of the problem of class overlap in imbalanced domains, essentially dividing the paper into two parts: a conceptual discussion of the problems (Sections 2 and 3) and a multi-view panorama of the current state of knowledge and open avenues across several fields of Machine Learning (Sections 4 to 6).

In the first part of the paper, acknowledging class overlap as the overarching problem (as per se it is more harmful than class imbalance), we start by discussing the concepts associated with its definition across related work. We reason towards the idea that class overlap comprises multiple sources of complexity and that it needs to be characterised accordingly. Indeed, we argue that the class overlap measures currently used in the literature are not representative of the class overlap problem as a whole, but that they rather provide an estimate of a specific type (representation) of class overlap.

[16] https://github.com/SCI2SUGR/KEEL
[17] https://www.cs.waikato.ac.nz/ml/weka/
[18] https://cran.r-project.org/web/packages/unbalanced
[19] https://cran.r-project.org/web/packages/smotefamily
[20] https://cran.r-project.org/web/packages/ROSE
[21] https://cran.r-project.org/web/packages/imbalance
[22] https://pypi.org/project/imbalanced-learn/
[23] https://github.com/analyticalmindsltd/smote_variants
[24] https://github.com/miriamspsantos/open-source-imbalance-overlap
[25] https://github.com/nmacia/dcol
[26] https://github.com/lpfgarcia/ECoL
[27] https://github.com/victorhb/ImbCoL
[28] https://github.com/lpfgarcia/SCoL
[29] https://github.com/rivolli/mfe
[30] https://github.com/ealcobaca/pymfe
[31] https://github.com/DiogoApostolo/pycol
[32] https://github.com/miriamspsantos/datagenerator
[33] https://matilda.unimelb.edu.au/matilda/our-methodology
[34] https://pypi.org/project/pyhard/

In this regard, in order to systematise the understanding of the problem of class overlap, we identify three main components underlying its characterisation: (1) the decomposition of the domains into regions of interest, (2) the identification of problematic regions (overlapped regions), and (3) the quantification/measurement of the class overlap problem. Depending on the approaches followed within each component, the obtained characterisation may refer to distinct class overlap representations, reflecting different insights on the problem.

Accordingly, we devise a novel taxonomy of class overlap complexity measures, establishing four main class overlap representations: (i) Feature Overlap, (ii) Instance Overlap, (iii) Structural Overlap, and (iv) Multiresolution Overlap. Each group is characterised in what concerns the insight its measures provide regarding the class overlap problem, as well as existing limitations. In other words, we explain how each group is able to capture a given representation of class overlap, while failing to perceive others. Besides establishing the association between complexity measures and their class overlap representations, our taxonomy evidences the core properties of the measures and provides an overview of the relationships between them. Additionally, it includes a comprehensive set of complexity measures, beyond the well-known measures initially proposed by Ho and Basu, and discusses whether they account for class imbalance, or how they can be extended to do so.

All in all, the concepts and ideas explored within the first part of this paper, culminating in the proposal of a new taxonomy of class overlap complexity measures, lay the foundation for a unified view of the problem of class overlap and may serve as a stepping stone for the design of improved measures and a characterisation of the problem as a whole in real-world domains.

Having laid out our conceptualisation of the problem of class overlap and its challenging aspects for imbalanced domains, we move towards the second part of the paper, offering a multi-view panorama regarding the synergy of both issues across four important areas of Machine Learning: Data Analysis, Data Preprocessing, Algorithm Design and Meta-learning. Regarding ongoing research directions, a few recent trends can be identified:

- A great amount of related work is currently focused on analysing the complexity of imbalanced classification tasks, either to establish the baseline difficulty of the learning process (data analysis) or to develop recommendation systems that compile this information and produce new inferences with various applications (meta-learning). Among existing data complexity measures, those associated to class overlap have provided the most perceptive insights. Nevertheless, due to the known biases introduced by the class imbalance problem, recent research is currently investigating adaptations of complexity measures to imbalanced domains, or focusing on the development of new measures that can take both issues simultaneously into account;
- Addressing multiple vortices of class overlap, i.e., considering distinct sources of complexity where can class overlap has synergetic effects (e.g., local, structural, density information), has proven to be a successful approach, both in the field of data preprocessing and regarding the development of specialised approaches. Simultaneously incorporating several sources of information into the solutions seems to be key to produce improved results, which endorses our understanding of class overlap as a heterogeneous concept with distinct representations, and shows that there is an advantage in considering their combination;
- Another emergent line of research is the creation of instance spaces where the class overlap problem can be assessed in a lower dimensional feature space, through data visualisation. This strategy resorts to dimensionality reduction techniques, where projections can be optimised in order to reveal linear trends between data complexity and classification performance.

Finally, we complemented the revision of the current state-of-the-art by incorporating our thoughts regarding several lines of research across the four identified areas of research. We consider the following to be the most pressing to consider in future work:

- The development of approaches to address other learning tasks beyond binary-classification problems. Most of existing work on class imbalance and overlap is devised for binary-classification domains, whereas the issues identified for other contexts (multi-class and singular problems) are yet to be faced;
- More extensive comparison of approaches to handle imbalanced and overlapped domains. In experimental studies, proposed methods are often evaluated against well-established approaches. New experiments should include emergent methods developed during most recent years. Additionally, a deeper characterisation of datasets and standardisation of performance metrics is necessary to guarantee representative testbeds and a fair comparison of approaches;
- Optimisation of hyperparameters for preprocessing and specialised approaches, based on the evaluation of data complexity measures. In imbalanced and overlapped contexts, hyperparameters are often defined according to heuristic solutions or tuned based on classification results. Although previous research in related fields (Meta-learning) has produced an interesting body of work on the topic of hyperparameter recommendation (although most often using traditional meta-features), further research on imbalanced and overlapped domains is required, and should explore the possibility of incorporating complexity measures into the tuning process;
- In addition to the previous point, despite the fact that the Deep Learning community has invested in addressing the class imbalance problem in the latest years, deep learning systems are rarely discussed in more challenging scenarios, namely those comprising additional difficult characteristics, such as class overlap. It would be important to strengthen the understanding we currently have on the behaviour or deep learning models, given that despite their growing interest in the machine learning community, they seem to suffer from the save handicaps as their classical counterparts, namely in what concerns the combination of class imbalance and overlap.
- More thorough studies on the effect of class imbalance and overlap on distinct learning biases. Existing studies comprise artificially generated data, with controlled parameters to create distinct complexity factors. New insights are needed for real-world domains;
- The creation of a comprehensive benchmark of datasets and their characterisation should also be prioritised in future research. The same applies to the development of open-source implementations of state-of-the-art approaches for imbalanced and overlapped domains, as well as data complexity measures beyond those established by Ho and Basu, which are mainly the focus of existing libraries.

In sum, the purpose and contribution of this manuscript is two-fold. First, it establishes the theoretical foundations of the problem of class overlap and its implications for imbalanced domains. It is our belief that, despite the increasing amount of proposals for new methods and approaches to address imbalanced and overlapped domains, the lack of understanding regarding the class overlap problem (i.e., the lack of a precise definition, measurement, and characterisation of the problem) is preventing the development of optimal solutions. In this regard, we hope that the concepts and resulting taxonomy discussed throughout this work, acknowledging the heterogeneity of the class overlap problem, may encourage the dialogue among researchers towards a consensus on the matter. Secondly, beyond providing a comprehensive identification of open avenues for research, this paper incorporates our thoughts and suggestions on how to address them in future work. We

sincerely hope that these lines of investigation may guide machine learning researchers on their journey to pursue future research in this field.

CRediT authorship contribution statement

Miriam Seoane Santos: Conceptualisation, Methodology, Literature Search, Investigation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Pedro Henriques Abreu:** Conceptualisation, Validation, Writing – review & editing, Supervision. **Nathalie Japkowicz:** Conceptualisation, Methodology, Formal analysis, Validation, Writing – review & editing. **Alberto Fernández:** Conceptualisation, Validation, Writing – original draft, Writing – review & editing, Supervision. **João Santos:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

This work is funded by the FCT - Foundation for Science and Technology, Portugal, I.P./MCTES through national funds (PIDDAC), within the scope of CISUC R&D Unit - UIDB/00326/2020 or project code UIDP/00326/2020. This work is also partially supported by Andalusian frontier regional project A-TIC-434-UGR20 and by the Spanish Ministry of Science and Technology under project PID2020-119478GB-I00 including European Regional Development Funds. The work is further supported by the FCT Research Grant, Portugal SFRH/BD/138749/2018.

References

[1] S. Das, S. Datta, B. Chaudhuri, Handling data irregularities in classification: Foundations, trends, and future challenges, Pattern Recognit. 81 (2018) 674–693.
[2] K. Napierała, J. Stefanowski, S. Wilk, Learning from imbalanced data in presence of noisy and borderline examples, in: International Conference on Rough Sets and Current Trends in Computing, Springer, 2010, pp. 158–167.
[3] V. López, A. Fernández, S. García, V. Palade, F. Herrera, An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics, Inform. Sci. 250 (2013) 113–141.
[4] J. Stefanowski, Dealing with data difficulty factors while learning from imbalanced data, in: Challenges in Computational Statistics and Data Mining, Springer, 2016, pp. 333–363.
[5] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, F. Herrera, Data intrinsic characteristics, Learn. Imbalanced Data Sets (2018) 253–277.
[6] S. Wojciechowski, S. Wilk, Difficulty factors and preprocessing in imbalanced data sets: An experimental study on artificial data, Found. Comput. Decis. Sci. 42 (2) (2017) 149–176.
[7] V. García, R. Mollineda, J. Sánchez, On the k-NN performance in a challenging scenario of imbalance and overlapping, Pattern Anal. Appl. 11 (3–4) (2008) 269–280.
[8] M.R. Smith, T. Martinez, C. Giraud-Carrier, An instance level analysis of data complexity, Mach. Learn. 95 (2) (2014) 225–256.
[9] A. Fernández, S. Garcia, F. Herrera, N.V. Chawla, SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary, J. Artificial Intelligence Res. 61 (2018) 863–905.
[10] M.S. Santos, J.P. Soares, P.H. Abreu, H. Araújo, J. Santos, Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches, IEEE Comput. Intell. Mag. 13 (3) (2018) 59–76.
[11] M. Denil, T. Trappenberg, Overlap versus imbalance, in: Canadian Conference on Artificial Intelligence, Springer, 2010, pp. 220–231.
[12] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, G. Bing, Learning from class-imbalanced data: Review of methods and applications, Expert Syst. Appl. 73 (2017) 220–239.
[13] H.K. Lee, S.B. Kim, An overlap-sensitive margin classifier for imbalanced and overlapping data, Expert Syst. Appl. 98 (2018) 72–83.
[14] R. Prati, B. G., M. Monard, Class imbalances versus class overlapping: An analysis of a learning system behavior, in: Mexican International Conference on Artificial Intelligence, Springer, 2004, pp. 312–321.
[15] M. Mercier, M.S. Santos, P.H. Abreu, C. Soares, J.P. Soares, J. Santos, Analysing the footprint of classifiers in overlapped and imbalanced contexts, in: International Symposium on Intelligent Data Analysis, Springer, 2018, pp. 200–212.
[16] G.-H. Fu, Y.-J. Wu, M.-J. Zong, L.-Z. Yi, Feature selection and classification by minimizing overlap degree for class-imbalanced data in metabolomics, Chemometr. Intell. Lab. Syst. 196 (2020) 103906.
[17] D. Singh, A. Gosain, A. Saha, Weighted k-nearest neighbor based data complexity metrics for imbalanced datasets, Stat. Anal. Data Min.: ASA Data Sci. J. 13 (4) (2020) 394–404.
[18] P. Vuttipittayamongkol, E. Elyan, A. Petrovski, On the class overlap problem in imbalanced data classification, Knowl.-Based Syst. (2020) 106631.
[19] M.S. Santos, P.H. Abreu, N. Japkowicz, A. Fernández, C. Soares, S. Wilk, J. Santos, On the joint-effect of class imbalance and overlap: A critical review, Artif. Intell. Rev. (2022) 1–69.
[20] T. Meng, X. Jing, Z. Yan, W. Pedrycz, A survey on machine learning for data fusion, Inf. Fusion 57 (2020) 115–129.
[21] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.
[22] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, J. Jorge, Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications, Inf. Fusion 81 (2022) 59–83.
[23] Y. Zhu, J. Ma, C. Yuan, X. Zhu, Interpretable learning based dynamic graph convolutional networks for alzheimer's disease analysis, Inf. Fusion 77 (2022) 53–61.
[24] J. Sun, H. Li, H. Fujita, B. Fu, W. Ai, Class-imbalanced dynamic financial distress prediction based on adaboost-SVM ensemble combined with SMOTE and time weighting, Inf. Fusion 54 (2020) 128–144.
[25] F. Ali, S. El-Sappagh, S.R. Islam, D. Kwak, A. Ali, M. Imran, K.-S. Kwak, A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion, Inf. Fusion 63 (2020) 208–222.
[26] Y. Zhang, S. Wang, K. Xia, Y. Jiang, P. Qian, A.D.N. Initiative, et al., Alzheimer's disease multiclass diagnosis via multimodal neuroimaging embedding feature selection and fusion, Inf. Fusion 66 (2021) 170–183.
[27] S.-H. Wang, D.R. Nayak, D.S. Guttery, X. Zhang, Y.-D. Zhang, COVID-19 classification by CCSHNet with deep fusion using transfer learning and discriminant correlation analysis, Inf. Fusion 68 (2021) 131–148.
[28] H. Yang, Y. Luo, X. Ren, M. Wu, X. He, B. Peng, K. Deng, D. Yan, H. Tang, H. Lin, Risk prediction of diabetes: big data mining with fusion of multifarious physical examination indicators, Inf. Fusion 75 (2021) 140–149.
[29] S.-H. Wang, V.V. Govindaraj, J.M. Górriz, X. Zhang, Y.-D. Zhang, Covid-19 classification by fgcnet with deep feature fusion from graph convolutional network and convolutional neural network, Inf. Fusion 67 (2021) 208–229.
[30] G. Muhammad, M.S. Hossain, COVID-19 and non-COVID-19 classification using multi-layers fusion from lung ultrasound images, Inf. Fusion 72 (2021) 80–88.
[31] L. Chen, B. Fang, Z. Shang, Y. Tang, Tackling class overlap and imbalance problems in software defect prediction, Softw. Qual. J. 26 (1) (2018) 97–125.
[32] M. Lopez-Martin, A. Sanchez-Esguevillas, J.I. Arribas, B. Carro, Supervised contrastive learning over prototype-label embeddings for network intrusion detection, Inf. Fusion 79 (2022) 200–228.
[33] T. Ho, M. Basu, Complexity measures of supervised classification problems, IEEE Trans. Pattern Anal. Mach. Intell. 24 (3) (2002) 289–300.
[34] N. Anwar, G. Jones, S. Ganesh, Measurement of data complexity for classification problems with unbalanced data, Stat. Anal. Data Min.: ASA Data Sci. J. 7 (3) (2014) 194–211.
[35] L. Cummins, Combining and choosing case base maintenance algorithms (Ph.D. thesis), University College Cork, 2013.
[36] E. Leyva, A. González, R. Perez, A set of complexity measures designed for applying meta-learning to instance selection, IEEE Trans. Knowl. Data Eng. 27 (2) (2014) 354–367.
[37] G. Armano, E. Tamponi, Experimenting multiresolution analysis for identifying regions of different classification complexity, Pattern Anal. Appl. 19 (1) (2016) 129–137.
[38] Z. Borsos, C. Lemnaru, R. Potolea, Dealing with overlap and imbalance: A new metric and approach, Pattern Anal. Appl. 21 (2) (2018) 381–395.
[39] A. Orriols-Puig, N. Macia, T.K. Ho, Documentation for the data complexity library in C++, Universitat Ramon Llull, la Salle 196 (2010) 1–40.
[40] A.C. Lorena, L.P. Garcia, J. Lehmann, M.C. Souto, T.K. Ho, How complex is your classification problem? A survey on measuring classification complexity, ACM Comput. Surv. 52 (5) (2019) 1–34.
[41] J.D. Pascual-Triana, D. Charte, M.A. Arroyo, A. Fernández, F. Herrera, Revisiting data complexity metrics based on morphology for overlap and imbalance: Snapshot, new overlap number of balls metrics and singular problems prospect, Knowl. Inf. Syst. (2021) 1–29.

[42] V.H. Barella, L.P. Garcia, M.C. de Souto, A.C. Lorena, A.C. de Carvalho, Assessing the data complexity of imbalanced datasets, Inform. Sci. 553 (2021) 83–109.

[43] A. Fernández, S. García, M. Galar, R.C. Prati, B. Krawczyk, F. Herrera, Learning from imbalanced data sets, Vol. 11, Springer, 2018.

[44] A. Rivolli, L.P. Garcia, C. Soares, J. Vanschoren, A.C. de Carvalho, Characterizing classification datasets: A study of meta-features for meta-learning, 2018, arXiv preprint arXiv:1808.10406.

[45] V. García, R. Alejo, J. Sánchez, J. Sotoca, R. Mollineda, Combined effects of class imbalance and class overlap on instance-based classification, in: International Conference on Intelligent Data Engineering and Automated Learning, Springer, 2006, pp. 371–378.

[46] V. García, R. Mollineda, J. Sánchez, R. Alejo, J. Sotoca, When overlapping unexpectedly alters the class imbalance effects, in: Iberian Conference on Pattern Recognition and Image Analysis, Springer, 2007, pp. 499–506.

[47] V. García, J. Sánchez, R. Mollineda, An empirical study of the behavior of classifiers on imbalanced and overlapped data sets, in: Iberoamerican Congress on Pattern Recognition, Springer, 2007, pp. 397–406.

[48] J. Stefanowski, Overlapping, rare examples and class decomposition in learning classifiers from imbalanced data, in: Emerging Paradigms in Machine Learning, Springer, 2013, pp. 277–306.

[49] X. Chen, L. Zhang, X. Wei, X. Lu, An effective method using clustering-based adaptive decomposition and editing-based diversified oversamping for multi-class imbalanced datasets, Appl. Intell. (2020) 1–16.

[50] Y. Zhu, Y. Yan, Y. Zhang, Y. Zhang, EHSO: Evolutionary hybrid sampling in overlapping scenarios for imbalanced learning, Neurocomputing 417 (2020) 333–346.

[51] J.M. Sotoca, J. Sanchez, R.A. Mollineda, A review of data complexity measures and their applicability to pattern classification problems, Actas Del III Taller Nacional de Mineria de Datos Y Aprendizaje. TAMIDA (2005) 77–83.

[52] J.M. Sotoca, R.A. Mollineda, J.S. Sánchez, A meta-learning framework for pattern classication by means of data complexity measures, Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial 10 (29) (2006) 31–38.

[53] J. Luengo, A. Fernández, S. García, F. Herrera, Addressing data complexity for imbalanced data sets: Analysis of SMOTE-based oversampling and evolutionary undersampling, Soft Comput. 15 (10) (2011) 1909–1936.

[54] V.H. Barella, L.P. Garcia, M.P. de Souto, A.C. Lorena, A. de Carvalho, Data complexity measures for imbalanced classification tasks, in: 2018 International Joint Conference on Neural Networks, IJCNN, IEEE, 2018, pp. 1–8.

[55] A. Ali, S.M. Shamsuddin, A.L. Ralescu, et al., Classification with class imbalance problem: A review, Int. J. Adv. Soft Comput. Appl. 7 (3) (2015) 176–204.

[56] C. M. Van der Walt, E. Barnard, Measures for the characterisation of pattern-recognition data sets, in: Annual Symposium of the Pattern Recognition Association of South Africa, 2007, pp. 1–6.

[57] J. Błaszczyński, J. Stefanowski, Local data characteristics in learning classifiers from imbalanced data, in: Advances in Data Analysis with Computational Intelligence Methods, Springer, 2018, pp. 51–85.

[58] S. Oh, A new dataset evaluation method based on category overlap, Comput. Biol. Med. 41 (2) (2011) 115–122.

[59] C. Thornton, Separability is a learner's best friend, in: 4th Neural Computation and Psychology Workshop, London, 9–11 April 1997, Springer, 1998, pp. 40–46.

[60] J. Greene, Feature subset selection using thornton's separability index and its applicability to a number of sparse proximity-based classifiers, in: Annual Symposium of the Pattern Recognition Association of South Africa, 2001, pp. 1–5.

[61] K. Napierala, J. Stefanowski, Types of minority class examples and their influence on learning classifiers from imbalanced data, J. Intell. Inf. Syst. 46 (3) (2016) 563–597.

[62] R.A. Sowah, M.A. Agebure, G.A. Mills, K.M. Koumadi, S.Y. Fiawoo, New cluster undersampling technique for class imbalance learning, Int. J. Mach. Learn. Comput. 6 (3) (2016) 205.

[63] A. Guzmán-Ponce, R.M. Valdovinos, J.S. Sánchez, J.R. Marcial-Romero, A new under-sampling method to face class overlap and imbalance, Appl. Sci. 10 (15) (2020) 5164.

[64] P. Vuttipittayamongkol, E. Elyan, Improved overlap-based undersampling for imbalanced dataset classification with application to epilepsy and parkinson's disease, Int. J. Neural Syst. 30 (08) (2020) 2050043.

[65] C.M. Van der Walt, et al., Data measures that characterise classification problems (Ph.D. thesis), University of Pretoria, 2008.

[66] S. Massie, S. Craw, N. Wiratunga, Complexity-guided case discovery for case based reasoning, in: AAAI, Vol. 5, 2005, pp. 216–221.

[67] S. Singh, PRISM–A novel framework for pattern recognition, Pattern Anal. Appl. 6 (2) (2003) 134–149.

[68] S. Singh, Multiresolution estimates of classification complexity, IEEE Trans. Pattern Anal. Mach. Intell. 25 (12) (2003) 1534–1539.

[69] C.G. Weng, J. Poon, A data complexity analysis on imbalanced datasets and an alternative imbalance recovering strategy, in: 2006 IEEE WIC ACM International Conference on Web Intelligence, IEEE, 2006, pp. 270–276.

[70] P. Vorraboot, S. Rasmequan, K. Chinnasarn, C. Lursinsap, Improving classification rate constrained to imbalanced data between overlapped and non-overlapped regions by hybrid algorithms, Neurocomputing 152 (2015) 429–443.

[71] E.R. Fernandes, A.C. de Carvalho, Evolutionary inversion of class distribution in overlapping areas for multi-class imbalanced learning, Inform. Sci. 494 (2019) 141–154.

[72] M. Lango, D. Brzezinski, J. Stefanowski, Imweights: Classifying imbalanced data using local and neighborhood information, in: Second International Workshop on Learning with Imbalanced Domains: Theory and Applications, PMLR, 2018, pp. 95–109.

[73] M. Lango, K. Napierala, J. Stefanowski, Evaluating difficulty of multi-class imbalanced data, in: International Symposium on Methodologies for Intelligent Systems, Springer, 2017, pp. 312–322.

[74] D. Charte, F. Charte, S. García, F. Herrera, A snapshot on nonstandard supervised learning problems: taxonomy, relationships, problem transformations and algorithm adaptations, Prog. Artif. Intell. 8 (1) (2019) 1–14.

[75] J.P.M. De Sá, Pattern Recognition: Concepts, Methods, and Applications, Springer Science & Business Media, 2001.

[76] C. Bunkhumpornpat, K. Sinapiromsaran, DBMUTE: density-based majority under-sampling technique, Knowl. Inf. Syst. 50 (3) (2017) 827–850.

[77] P. Vuttipittayamongkol, E. Elyan, A. Petrovski, C. Jayne, Overlap-based undersampling for improving imbalanced data classification, in: International Conference on Intelligent Data Engineering and Automated Learning, Springer, 2018, pp. 689–697.

[78] C. Bunkhumpornpat, K. Sinapiromsaran, C. Lursinsap, MUTE: Majority undersampling technique, in: 2011 8th International Conference on Information, Communications & Signal Processing, IEEE, 2011, pp. 1–4.

[79] P. Vuttipittayamongkol, E. Elyan, Neighbourhood-based undersampling approach for handling imbalanced and overlapped data, Inform. Sci. 509 (2020) 47–70.

[80] J. Sáez, J. Luengo, J. Stefanowski, F. Herrera, SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering, Inform. Sci. 291 (2015) 184–203.

[81] I. Nekooeimehr, S.K. Lai-Yuen, Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets, Expert Syst. Appl. 46 (2016) 405–416.

[82] J. Wei, H. Huang, L. Yao, Y. Hu, Q. Fan, D. Huang, IA-SUWO: An improving adaptive semi-unsupervised weighted oversampling for imbalanced classification problems, Knowl.-Based Syst. 203 (2020) 106116.

[83] J. Wei, H. Huang, L. Yao, Y. Hu, Q. Fan, D. Huang, NI-MWMOTE: An improving noise-immunity majority weighted minority oversampling technique for imbalanced classification problems, Expert Syst. Appl. 158 (2020) 113504.

[84] T. Zhu, Y. Lin, Y. Liu, Improving interpolation-based oversampling for imbalanced data learning, Knowl.-Based Syst. 187 (2020) 104826.

[85] G. Douzas, F. Bacao, Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE, Inform. Sci. 501 (2019) 118–135.

[86] V. García, J. Sánchez, A. Marqués, R. Florencia, G. Rivera, Understanding the apparent superiority of over-sampling through an analysis of local information for class-imbalanced data, Expert Syst. Appl. 158 (2020) 113026.

[87] A.R.S. Parmezan, H.D. Lee, F.C. Wu, Metalearning for choosing feature selection algorithms in data mining: Proposal of a new framework, Expert Syst. Appl. 75 (2017) 1–24.

[88] L.C. Okimoto, R.M. Savii, A.C. Lorena, Complexity measures effectiveness in feature selection, in: 2017 Brazilian Conference on Intelligent Systems, BRACIS, IEEE, 2017, pp. 91–96.

[89] L.C. Okimoto, A.C. Lorena, Data complexity measures in feature selection, in: 2019 International Joint Conference on Neural Networks, IJCNN, IEEE, 2019, pp. 1–8.

[90] B. Seijo-Pardo, V. Bolón-Canedo, A. Alonso-Betanzos, On developing an automatic threshold applied to feature selection ensembles, Inf. Fusion 45 (2019) 227–245.

[91] N.T. Dong, M. Khosla, Revisiting feature selection with data complexity, in: 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering, BIBE, IEEE, 2020, pp. 211–216.

[92] A. Fernández, M.J. del Jesus, F. Herrera, Addressing overlapping in classification with imbalanced datasets: A first multi-objective approach for feature and instance selection, in: International Conference on Intelligent Data Engineering and Automated Learning, Springer, 2015, pp. 36–44.

[93] X. Lin, H. Song, M. Fan, W. Ren, L. Li, W. Yao, The feature selection algorithm based on feature overlapping and group overlapping, in: 2016 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2016, pp. 619–624.

[94] H. Hartono, E. Ongko, Y. Risyani, Combining feature selection and hybrid approach redefinition in handling class imbalance and overlapping for multi-class imbalanced, Indonesian J. Electr. Eng. Comput. Sci. 21 (3) (2021) 1513–1522.

[95] B. Omar, F. Rustam, A. Mehmood, G.S. Choi, et al., Minimizing the overlapping degree to improve class-imbalanced learning under sparse feature selection: Application to fraud detection, IEEE Access 9 (2021) 28101–28110.

[96] K. Smith-Miles, D. Baatar, B. Wreford, R. Lewis, Towards objective measures of algorithm performance across instance space, Comput. Oper. Res. 45 (2014) 12–24.

[97] K. Smith-Miles, T.T. Tan, Measuring algorithm footprints in instance space, in: 2012 IEEE Congress on Evolutionary Computation, IEEE, 2012, pp. 1–8.

[98] M.A. Muñoz, L. Villanova, D. Baatar, K. Smith-Miles, Instance spaces for machine learning classification, Mach. Learn. 107 (1) (2018) 109–147.

[99] M.A. Muñoz, T. Yan, M.R. Leal, K. Smith-Miles, A.C. Lorena, G.L. Pappa, R.M. Rodrigues, An instance space analysis of regression problems, ACM Trans. Knowl. Discov. Data (TKDD) 15 (2) (2021) 1–25.

[100] J. Vanschoren, Meta-learning: A survey, 2018, arXiv preprint arXiv:1810.03548.

[101] M.M. Nwe, K.T. Lynn, Knn-based overlapping samples filter approach for classification of imbalanced data, in: International Conference on Software Engineering Research, Management and Applications, Springer, 2019, pp. 55–73.

[102] P. Skryjomski, B. Krawczyk, Influence of minority class instance types on SMOTE imbalanced data oversampling, in: First International Workshop on Learning with Imbalanced Domains: Theory and Applications, PMLR, 2017, pp. 7–21.

[103] J. Sáez, B. Krawczyk, M. Woźniak, Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets, Pattern Recognit. 57 (2016) 164–178.

[104] M. Koziarski, M. Woźniak, CCR: A combined cleaning and resampling algorithm for imbalanced data classification, Int. J. Appl. Math. Comput. Sci. 27 (4) (2017) 727–736.

[105] A. Fernández, C.J. Carmona, M. José del Jesus, F. Herrera, A Pareto-based ensemble with feature and instance selection for learning from multi-class imbalanced datasets, Int. J. Neural Syst. 27 (06) (2017) 1750028.

[106] V. H. Barella, E. P. Costa, A. C.P.L.F. de Carvalho, Clusteross: A new undersampling method for imbalanced learning, in: Brazilian Conference on Intelligent Systems, Academic Press, 2014, pp. 1–6.

[107] K. Ghosh, C. Bellinger, R. Corizzo, B. Krawczyk, N. Japkowicz, On the combined effect of class imbalance and concept complexity in deep learning, 2021, arXiv preprint arXiv:2107.14194.

[108] A. Rivolli, L.P. Garcia, C. Soares, J. Vanschoren, A.C. de Carvalho, Towards reproducible empirical research in meta-learning, 2018, pp. 32–52, arXiv preprint arXiv:1808.10406.

[109] S.N. das Dôres, L. Alves, D.D. Ruiz, R.C. Barros, A meta-learning framework for algorithm recommendation in software fault prediction, in: Proceedings of the 31st Annual ACM Symposium on Applied Computing, 2016, pp. 1486–1491.

[110] R. Shah, V. Khemani, M. Azarian, M. Pecht, Y. Su, Analyzing data complexity using metafeatures for classification algorithm selection, in: 2018 Prognostics and System Health Management Conference (PHM-Chongqing), IEEE, 2018, pp. 1280–1284.

[111] X. Zhang, R. Li, B. Zhang, Y. Yang, J. Guo, X. Ji, An instance-based learning recommendation algorithm of imbalance handling methods, Appl. Math. Comput. 351 (2019) 204–218.

[112] A.J. Costa, M.S. Santos, C. Soares, P.H. Abreu, Analysis of imbalance strategies recommendation using a meta-learning approach, in: 7th ICML Workshop on Automated Machine Learning (AutoML-ICML2020), 2020, pp. 1–10.

[113] L.P. Garcia, A.C. Lorena, M.C. de Souto, T.K. Ho, Classifier recommendation using data complexity measures, in: 2018 24th International Conference on Pattern Recognition, ICPR, IEEE, 2018, pp. 874–879.

[114] J. Luengo, F. Herrera, An automatic extraction method of the domains of competence for learning classifiers using data complexity measures, Knowl. Inf. Syst. 42 (1) (2015) 147–180.

[115] Z. Liu, W. Cao, Z. Gao, J. Bian, H. Chen, Y. Chang, T.-Y. Liu, Self-paced ensemble for highly imbalanced massive data classification, in: 2020 IEEE 36th International Conference on Data Engineering, ICDE, IEEE, 2020, pp. 841–852.

[116] J.A. Sáez, M. Galar, B. Krawczyk, Addressing the overlapping data problem in classification using the one-vs-one decomposition strategy, IEEE Access 7 (2019) 83396–83411.

[117] M. Galar, A. Fernández, E. Barrenechea, F. Herrera, DRCW-OVO: distance-based relative competence weighting combination for one-vs-one strategy in multi-class problems, Pattern Recognit. 48 (1) (2015) 28–42.

[118] M. Janicka, M. Lango, J. Stefanowski, Using information on class interrelations to improve classification of multiclass imbalanced data: A new resampling algorithm, Int. J. Appl. Math. Comput. Sci. 29 (4) (2019).

[119] F. Herrera, F. Charte, A.J. Rivera, M.J. Del Jesus, Multilabel classification, in: Multilabel Classification, Springer, 2016, pp. 17–31.

[120] I. Bendjoudi, F. Vanderhaegen, D. Hamad, F. Dornaika, Multi-label, multi-task CNN approach for context-based emotion recognition, Inf. Fusion 76 (2021) 422–428.

[121] F. Herrera, S. Ventura, R. Bello, C. Cornelis, A. Zafra, D. Sánchez-Tarragó, S. Vluymans, Multiple instance learning, in: Multiple Instance Learning, Springer, 2016, pp. 17–33.

[122] S. Vluymans, D.S. Tarragó, Y. Saeys, C. Cornelis, F. Herrera, Fuzzy rough classifiers for class imbalanced multi-instance data, Pattern Recognit. 53 (2016) 36–45.

[123] G. Melki, A. Cano, S. Ventura, MIRSVM: multi-instance support vector machine with bag representatives, Pattern Recognit. 79 (2018) 228–241.

[124] S. Sun, L. Mao, Z. Dong, L. Wu, Multiview Machine Learning, Springer, 2019.

[125] D. Jiang, R. Xu, X. Xu, Y. Xie, Multi-view feature transfer for click-through rate prediction, Inform. Sci. 546 (2021) 961–976.

[126] R.G. Mantovani, A.L. Rossi, J. Vanschoren, B. Bischl, A.C. Carvalho, To tune or not to tune: recommending when to adjust SVM hyper-parameters via meta-learning, in: 2015 International Joint Conference on Neural Networks, IJCNN, IEEE, 2015, pp. 1–8.

[127] R.G. Mantovani, A.L. Rossi, J. Vanschoren, A.C. de Carvalho, Meta-learning recommendation of default hyper-parameter values for SVMs in classification tasks, in: MetaSel PKDD/ECML, 2015, pp. 80–92.

[128] M. Mahin, M.J. Islam, B.C. Debnath, A. Khatun, Tuning distance metrics and k to find sub-categories of minority class from imbalance data using k nearest neighbours, in: 2019 International Conference on Electrical, Computer and Communication Engineering, ECCE, IEEE, 2019, pp. 1–6.

[129] N. Macià, E. Bernadó-Mansilla, Towards UCI+: A mindful repository design, Inform. Sci. 261 (2014) 237–262.

[130] L.P. Garcia, A. Rivolli, E. Alcobaça, A.C. Lorena, A.C. de Carvalho, Boosting meta-learning with simulated data complexity measures, Intell. Data Anal. 24 (5) (2020) 1011–1028.

[131] V.V. de Melo, A.C. Lorena, Using complexity measures to evolve synthetic classification datasets, in: 2018 International Joint Conference on Neural Networks, IJCNN, IEEE, 2018, pp. 1–8.

[132] A. Correia, C. Soares, A. Jorge, Dataset morphing to analyze the performance of collaborative filtering, in: International Conference on Discovery Science, Springer, 2019, pp. 29–39.

[133] T.R. França, P.B. Miranda, R.B. Prudêncio, A.C. Lorenaz, A.C. Nascimento, A many-objective optimization approach for complexity-based data set generation, in: 2020 IEEE Congress on Evolutionary Computation, CEC, IEEE, 2020, pp. 1–8.

[134] J. Alcalá-Fdez, L. Sánchez, S. Garcia, M.J. del Jesus, S. Ventura, J.M. Garrell, J. Otero, C. Romero, J. Bacardit, V.M. Rivas, et al., KEEL: A software tool to assess evolutionary algorithms for data mining problems, Soft Comput. 13 (3) (2009) 307–318.

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

[136] I. Triguero, S. González, J.M. Moyano, S. García, J. Alcalá-Fdez, J. Luengo, A. Fernández, M.J. del Jesús, L. Sánchez, F. Herrera, KEEL 3.0: An open source software for multi-stage analysis in data mining, Int. J. Comput. Intell. Syst. 10 (1) (2017) 1238–1249.

[137] E. Frank, M. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I.H. Witten, L. Trigg, Weka-a machine learning workbench for data mining, in: Data Mining and Knowledge Discovery Handbook, Springer, 2009, pp. 1269–1277.

[138] A. Dal Pozzolo, O. Caelen, S. Waterschoot, G. Bontempi, Racing for unbalanced methods selection, in: International Conference on Intelligent Data Engineering and Automated Learning, Springer, 2013, pp. 24–31.

[139] N. Lunardon, G. Menardi, N. Torelli, ROSE: A package for binary imbalanced learning, R Journal 6 (1) (2014).

[140] W. Siriseriwan, Smotefamily: A collection of oversampling techniques for class imbalance problem based on SMOTE, 2019.

[141] I. Cordón, S. García, A. Fernández, F. Herrera, Imbalance: Oversampling algorithms for imbalanced classification in R, Knowl.-Based Syst. 161 (2018) 329–341.

[142] G. Lemaître, F. Nogueira, C.K. Aridas, Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning, J. Mach. Learn. Res. 18 (1) (2017) 559–563.

[143] G. Kovács, An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets, Appl. Soft Comput. 83 (2019) 105662.

[144] E. Alcobaça, F. Siqueira, A. Rivolli, L.P.F. Garcia, J.T. Oliva, A.C.P.L.F. de Carvalho, MFE: Towards reproducible meta-feature extraction, J. Mach. Learn. Res. 21 (111) (2020) 1–5.

[145] P.Y.A. Paiva, K. Smith-Miles, M.G. Valeriano, A.C. Lorena, Pyhard: A novel tool for generating hardness embeddings to support data-centric analysis, 2021, arXiv preprint arXiv:2109.14430.