Noname manuscript No.
(will be inserted by the editor)

On the joint-effect of Class Imbalance and Overlap

A Critical Review

Miriam Seoane Santos · Pedro Henriques Abreu · Nathalie Japkowicz · Alberto Fernández · Carlos Soares · Szymon Wilk · João Santos

Received: date / Accepted: date

Abstract Current research on imbalanced data recognises that class imbalance is aggravated by other data intrinsic characteristics, among which class overlap stands out as one of the most harmful. The combination of these two problems creates a new and difficult scenario for classification tasks and has been discussed in several research works over the past two decades. In this paper, we argue that despite some insightful information can be derived from related research, the joint-effect of class overlap and imbalance is still not fully understood, and advocate for the need to move towards a unified view of the class overlap problem in imbalanced domains. To that end, we start by performing a thorough analysis of existing literature on the joint-effect of class imbalance and overlap, elaborating on important details left undiscussed on the original papers, namely the impact of data domains with different characteristics and the behaviour of classifiers with distinct learning

Miriam Seoane Santos and Pedro Henriques Abreu
University of Coimbra, Centre for Informatics and Systems of the University of Coimbra, Department of Informatics Engineering, Coimbra, Portugal
E-mail: miriams@dei.uc.pt, pha@dei.uc.pt

Nathalie Japkowicz
Department of Computer Science, American University, Washington, DC, 20016, USA
E-mail: nathalie.japkowicz@american.edu

Alberto Fernández
Department of Computer Science and Artificial Intelligence, University of Granada, Spain
E-mail: alberto@decsai.ugr.es

Carlos Soares
Fraunhofer Portugal AICOS and LIACC, Faculdade de Engenharia, Universidade do Porto, Porto, Portugal
E-mail: csoares@fe.up.pt

Szymon Wilk
Institute of Computing Science, Poznan University of Technology, Poznan, Poland
E-mail: szymon.wilk@cs.put.poznan.pl

João Santos
IPO-Porto Research Centre (CI-IPOP), Porto, Portugal
Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto, Portugal
E-mail: joao.santos@ipoporto.min-saude.pt

biases. This leads to the hypothesis that class overlap comprises multiple representations, which are important to accurately measure and analyse in order to provide a full characterisation of the problem. Accordingly, we devise two novel taxonomies, one for class overlap measures and the other for class overlap-based approaches, both resonating with the distinct representations of class overlap identified. This paper therefore presents a global and unique view on the joint-effect of class imbalance and overlap, from precursor work to recent developments in the field. It meticulously discusses some concepts taken as implicit in previous research, explores new perspectives in light of the limitations found, and presents new ideas that will hopefully inspire researchers to move towards a unified view on the problem and the development of suitable strategies for imbalanced and overlapped domains.

Keywords Class Imbalance · Class Overlap · Data Intrinsic Characteristics · Class Overlap Complexity Measures · Class Overlap-Based Approaches · Class Overlap Representations

1 Introduction

Class imbalance refers to a disproportion in the number of examples belonging to each class of a dataset and is known to bias classifiers towards the most represented concepts [43]. This situation is especially critical when minority class concepts are associated with a higher misclassification cost, such as the diagnosis of rare diseases [110, 115]. Although this is an important problem in isolation, its combination with other factors creates a much more difficult setting for classifiers, as growing research has brought to light [86, 101, 125]. These are referred to as *data intrinsic characteristics* [43, 86], *data difficulty factors* [125, 145] or *data irregularities* [34], and among others, include the problem of class overlap.

Class overlap has received much attention in the past two decades, since it is a source of complexity for traditional classification paradigms (e.g., max-margin classifiers, Bayesian classifiers, decision trees) [34, 62] and has been observed in several application domains (e.g., character recognition [84], software defect prediction [23] and protein and drug discovery [90, 114]). Indeed, among all data intrinsic characteristics, class overlap has been recognised as the most harmful issue for pattern classification [47, 58, 120] and remains one of the most studied topics nowadays [51, 116, 135]. Assuming an equal representation of classes (i.e., balanced domains), class overlap occurs when regions of the data space are populated by a similar number of training examples of each class [86, 36, 81]: as classes are equally represented in the same regions, their discrimination becomes more complicated. In imbalanced domains, the problem is aggravated since the few minority examples that exist may be mostly located in regions populated by the other class(es) as well.

Over the years, several research works have focused on characterising the combined effects of class imbalance and overlap. To that end, researchers created several synthetic data domains with different imbalance ratios and overlap degrees. Then, one or several classifiers were tested and classification results were evaluated, showing that class imbalance alone cannot be responsible for the deterioration of classification performance, and that class overlap plays an important role as well.

Therefore, the focus of related work was, essentially, to establish class overlap as a difficulty factor for classification tasks, especially in the presence of class imbalance. That caused the analysis of other important aspects to be neglected to some extent, such as the learning biases of used classifiers and the peculiarities of the considered data domains. In fact, some authors consider only a single classifier [55, 36, 106] or similar learning paradigms (e.g., tree and rule-based classifiers) [101], while the data domains are also considerably different among research works. By cross-referencing the obtained results across related work, important aspects that remained vague or understudied in previous research can now be brought to discussion on a deeper level.

In this work, we review the existing literature on the joint-effect of class imbalance and overlap, summarising their main conclusions and performing a thorough cross-referencing of results in order to analyse some details left undiscussed in the original papers. In particular, we focus on analysing the effect of the characteristics of studied data domains (e.g., data decomposition, structure, dimensionality and data typology) and the behaviour of classifiers with distinct biases (instance-based, rule and tree-based, Bayesian classifiers, neural networks, support vector machines and linear discriminants). A cross-reference of research results allows the evaluation of classifiers under several conditions (data domains, dimensionality, class imbalance and overlap) and effects on classification performance are explained from a theoretical (considering the known biases of classifiers) and empirical (considering the used data domains and obtained experimental results) perspective. In sum, we extend the current body of knowledge on the combination of class imbalance and overlap by focusing on the following research topics:

- What is the influence of intrinsic data characteristics (data decomposition, data structure, data dimensionality, data typology) on the classification performance for imbalanced and overlapped domains?
- How do classifiers with different nature (distinct learning biases) handle imbalanced and overlapped domains?

The analysis conducted over seminal work yielded important insights regarding the joint-effect of class imbalance and overlap. First, it allowed to derive some important lessons learned regarding the characteristics of the domains and nature of classifiers, two understudied topics that remained mostly hidden in related research. Then, it allowed to identify important limitations regarding the characterisation of class overlap in imbalanced domains and ultimately, to the idea that class overlap comprises several representations, which need to be quantified and analysed accordingly. On that note, a discussion on identifiability and quantification of class overlap, especially in real-world domains, arises naturally. We therefore provide a comprehensive review of class overlap measures and establish a novel taxonomy that defines distinct groups of measures according to the class overlap representations they are able to characterise. We conclude the paper by analysing emergent class overlap-based approaches applied to real-world imbalanced domains. It is our intent to show that, despite recent work suffers from the same limitations found in seminal work in what concerns the characterisation and quantification of class overlap, it is possible to associate the underlying behaviour of approaches to the class overlap representations they are attentive to. Establishing this association is a step towards the choice and development of specialised approaches depending on the characteristics of the domains. We therefore devise a

taxonomy of class overlap-based approaches aligned with the taxonomy proposed for class overlap measures.

Existing surveys mostly provide a bird's eye view on handling imbalanced data classification, state-of-the-art methods and applications, and current trends [62, 72, 76], although setting aside the study of other difficulty factors embedded in the nature of data. Some also touch upon the definition of data characteristics and their impact on classification tasks [34, 46]; however, without a specific focus on the joint-effect of class imbalance and overlap and its synergy with other characteristics of the data domains, different learning biases, quantification, or contemporary approaches. Related research in the field of classification complexity provides a generic overview of data complexity measures and their use across several application areas [88]. However, there is no established set of complexity measures for class overlap, as measures are grouped according to their underlying quantification mechanisms (e.g., feature-based, neighbourhood-based), rather than the insight they provide on the domain (e.g., feature overlap, instance overlap, structural overlap). Several recent measures linked to the class overlap problem are also comprised in an extra-category instead of thoroughly reviewed, as the main complexity measures described refer to those proposed by Ho and Basu on their pioneer work on the topic [66]. There is also no discussion in what concerns the adaptation of existing measures to imbalanced domains. The most related research is perhaps the recent review by Pattaramon et al. [135], which also discusses some emergent class overlap-based methods in imbalanced domains. However, no considerations regarding a taxonomy of methods or representations of class overlap are given. Of note is that authors also agree with the need of a well-established definition and measurement of class overlap and a standard measure for the class overlap degree in real-world domains, meeting our line of thought. What we put forward with this research is precisely a first step towards a consensus of the research community on this matter.

Contrary to previous works, this paper focuses on a critical analysis on the problem of class overlap in imbalanced domains, as it is considered the most harmful issue among data difficulty factors [47]. It focus specifically on the interrelations of class imbalance and overlap and their joint-effect on classification performance, considering two other influential factors often neglected in related research: the characteristics of the data domains, and the nature of classifiers. Additionally, more than providing a comprehensive review of related work in the field, this work presents an in-depth conceptual discussion of key concepts, scrutinising some of the assumptions and insights from previous work and their implications for real-world domains. What follows is a conceptualisation of class overlap as a heterogeneous concept, comprising multiple sources of complexity, and a theoretical evaluation of its challenging aspects for imbalanced domains. We also present a critical discussion on both *i)* the identifiability and quantification of class overlap in real-world contexts and *ii)* the state-of-the-art methods to handle the problem in imbalanced data contexts.

In sum, we provide a global and unique vision on the joint-effect of class imbalance and overlap, identifying existing theoretical and empirical limitations in previous and current research, and discussing new ideas that advocate towards a unified view on the problem. In detail, the contributions of this work are as follows: (i) a revision of related work on the joint-effect of class imbalance and overlap; (ii) a discussion of the impact of intrinsic data characteristics in syn-

ergy with class imbalance and overlap; (iii) an overview of the joint-effect of class overlap and imbalance on the performance of classifiers with different learning biases; (iv) a motivation for the characterisation of class overlap according to different perspectives and a discussion of distinct class overlap representations; (v) a review of measures of class overlap and a taxonomy aligned with its different representations; (vi) a review of the state-of-the-art approaches for imbalanced and overlapped domains and a taxonomy that resonates with the identified class overlap representations; and (vii) the identification of limitations of previous and current research and a motivation for a unified view of the class overlap problem in imbalanced domains.

To our knowledge, this work provides the most comprehensive review on the subject, from seminal work to emergent research. More importantly, this is the first work to put forward that class overlap observes a multitude of representations and systematises both class overlap measures and approaches towards that characterisation.

The reader should navigate this paper as follows. Section 2 reviews seminal work on class imbalance and overlap, describing the experiments and data domains in detail and elaborating on their main conclusions. Then, Sections 3 and 4 discuss the lessons learned with respect to the impact of the characteristics of the data domains and the learning biases of distinct classifiers, respectively. While Section 3 hints at distinct representations of class overlap, Section 4 reinforces the idea that linking the behaviour of classifiers to the characterisation of domains would prove transformative to future research in the field. In Section 5, we detail the limitations found in seminal work on synthetic data and discuss why they prevent a full understanding of the joint-effect of class overlap and imbalance, while also motivating the need to revise existing solutions for real-world domains. Hence, Sections 6 and 7 are focused on revising class overlap measures and class overlap-based approaches applied to real-world imbalanced domains. We start both sections by presenting a global view on the topic and introducing our proposed taxonomies with supporting schemas. Then, class overlap measures are described, formalised, and illustrated in detail, and class overlap-based approaches are presented, respectively, both divided by category. At the end of each section we present our summarising comments, discussing the most important limitations and open challenges for research. Finally, Section 8 summarises future directions that the research community should debate for a renewed view on the joint-effect of class overlap and imbalance, hopefully leading to new breakthroughs in the field, whereas Section 9 ends the paper, providing an overview of the main topics discussed throughout this work.

2 On the joint-effect of Class Imbalance and Overlap

In this section, we review the existing literature on the joint-effect of class imbalance and overlap. To help the reader navigate this section, Table 1 presents the related work in chronological order, focusing on their objectives, characterisation of data domains, experimental design (controlled parameters and studied classifiers), and main conclusions. In what follows, we discuss the related research, showing how the co-occurrence of class imbalance and overlap poses a more difficult problem that solving each issue independently. We focus on the global insights

regarding the joint-effect of class imbalance and overlap rather than the details of each research work. In Sections 3 and 4, we will elaborate on the lessons learned in what concerns the characteristics of the studied domains and classifiers.

Prati et al. [106] experimented with several variations of class imbalance and overlap by studying two Gaussian clusters where the distribution of minority and majority examples, as well as the distance between cluster centroids, could be changed (Figure 1). Authors showed that when the distance between class centroids was zero, the classification was extremely difficult, independently of the considered class imbalance. Conversely, as the distance between class centroids increased, the class overlap problem ceased to exist and the classification results were high, independently of the percentage of minority examples.

García et al. [55] studied the combined effects of these two problems on instance-based classification algorithms (1-nearest neighbour classifier). Authors used artificial domains composed of two squares, each having a uniform distribution of points from the majority and minority classes, respectively (Figure 1). Whereas the class imbalance was fixed, the class overlap was manipulated through the distance between square centres, i.e., the majority class was moved towards the minority class in a stepwise manner (as per the original paper, we will refer to this configuration as a "typical situation"). While the classification results were maximal when there was no class overlap, the performance degraded as the overlap increased.

In [56, 57], in addition to typical situations, García et al. focused on a particular imbalanced scenario where the minority class was more represented than the majority class in the overlap region (considered an "atypical situation", as shown in Figure 1). In this case, the local class imbalance (in the overlap region) was different from the global class imbalance (in the entire domain). Authors considered several classification paradigms (please refer to Table 1) and showed that in typical situations, the classification performance of all classifiers on the minority class degraded with increasing class overlap. However, local classifiers were more suited to the recognition of the minority class, while global classifiers performed better on the majority class. In atypical situations, classifiers with a global nature benefited the recognition of the minority class, while local classifiers were better for the majority class.

In [58], García et al. further focused on the performance of KNN classifier (varying the value of k) versus the performance of other classifiers (Table 1) in typical and atypical situations, aiming to explain the influence of overall imbalance, local imbalance and the size of the overlap region on the behaviour of KNN classifier. In typical situations, smaller values of k were more suited to the recognition of the minority class, whereas higher values benefited the recognition of majority class examples. In turn, for atypical situations, the increase of k benefited the minority class and no significant changes occurred in the performance of the majority class, showing that KNN was more dependent on the local imbalance than on the global imbalance. When the overlap region was not balanced, the local imbalance ratio was more important than the size of the overlap region for KNN performance. Finally, for similar configurations of class imbalance and overlap, authors found that the complexity of the boundary decision was yet another difficulty factor for classifiers [58].

Denil and Trappenberg [36] studied the joint-effect of class imbalance and overlap on the performance of Support Vector Machines (SVM) by varying factors

individually and simultaneously for different training set sizes (Figure 1). For small training set sizes, as well as for small amounts of overlap and imbalance, the performance of SVM assuming that these factors are independent was similar to the one obtained from their combination. As the training set size increased, the influence of class imbalance was negligible and class overlap was the main responsible for the performance degradation. Thus, assuming that both factors were independent, the performance results obtained for large training sets in the presence of overlap alone should have been similar to the performance when both factors were present in data. However, the performance was even lower, indicating that the issues were far more serious in combination that in isolation [36].

Related work on the joint-impact of class overlap and imbalance also includes the research of Napierala et al. [101], Stefanowski [124], and Wojciechowski and Wilk [145]. Rather than considering overlap regions or areas, the focus shifted to the data typology of the minority class (i.e., considering different types of data examples) to approximate certain difficulty factors, such as class overlap. Class overlap was approximated by focusing on *borderline* examples, as they are highly related to the problem of class overlap (i.e., they appear in the borderline between classes). Overall, authors studied the influence of disturbing the minority class boundaries by adding an increasing number of borderline examples to domains with different characteristics – *paw*, *clover/flower* and *subclus* domains (Figure 1). Napierala et al. [101] showed that increasing the number of borderline examples highly degraded the performance of classifiers. Stefanowski [124, 125] focused on the *subclus* dataset and studied the impact of changing the number of subclusters (class decomposition), changing the percentage of borderline minority examples (class overlap) and changing the imbalance ratio. Experiments showed that the combination of class decomposition and overlap seemed to affect classification performance more than the increase of the imbalance ratio, and that for non-linear shapes the performance degradation was more accentuated. Wojciechowski and Wilk [145] further showed that data typology significantly affected the classification results more than class imbalance or data dimensionality.

Finally, Mercier et al. [98] reproduced several artificial data domains considered in previous works and analysed the performance degradation of classifiers with different learning biases (please refer to Table 1). Classifiers that learn on the basis of data space fragmentation were less affected by class overlap than linear classifiers (further details will be given throughout Section 4).

According to the key insights of the discussed research, the following conclusions can be established:

- Class overlap acts as a difficulty factor for classification, more than class imbalance. Indeed, although the class imbalance generally deteriorates the performance of classifiers, if there are no other complex data characteristics, then the class imbalance itself does not affect classification, regardless of the imbalance ratio [106, 55];
- These two problems do not have independent effects and the degradation caused by their combination is not equivalent to the aggregation of the degradation caused by each one individually [36]. Class overlap and imbalance have hidden dependencies that are not noticeable by analysing them separately;
- The joint-effect of class imbalance and overlap strongly depends on the nature of classifiers, the general characteristics of the domain (class decomposition,

data dimensionality, complexity of the decision boundaries) and on the local characteristics of the overlap region (local imbalance and data typology) [58, 98, 145].

In the following sections, we will detail the lessons learned in what concerns the characteristics of the studied domains and classifiers. The provided analysis is supported by a thorough examination of experimental results obtained in related research, which were aggregated by data domain and classifier[1].

[1] The reader may find supporting information in the supplementary material online at https://student.dei.uc.pt/~miriams/pdf-files/AIR_2021_Appendix.pdf

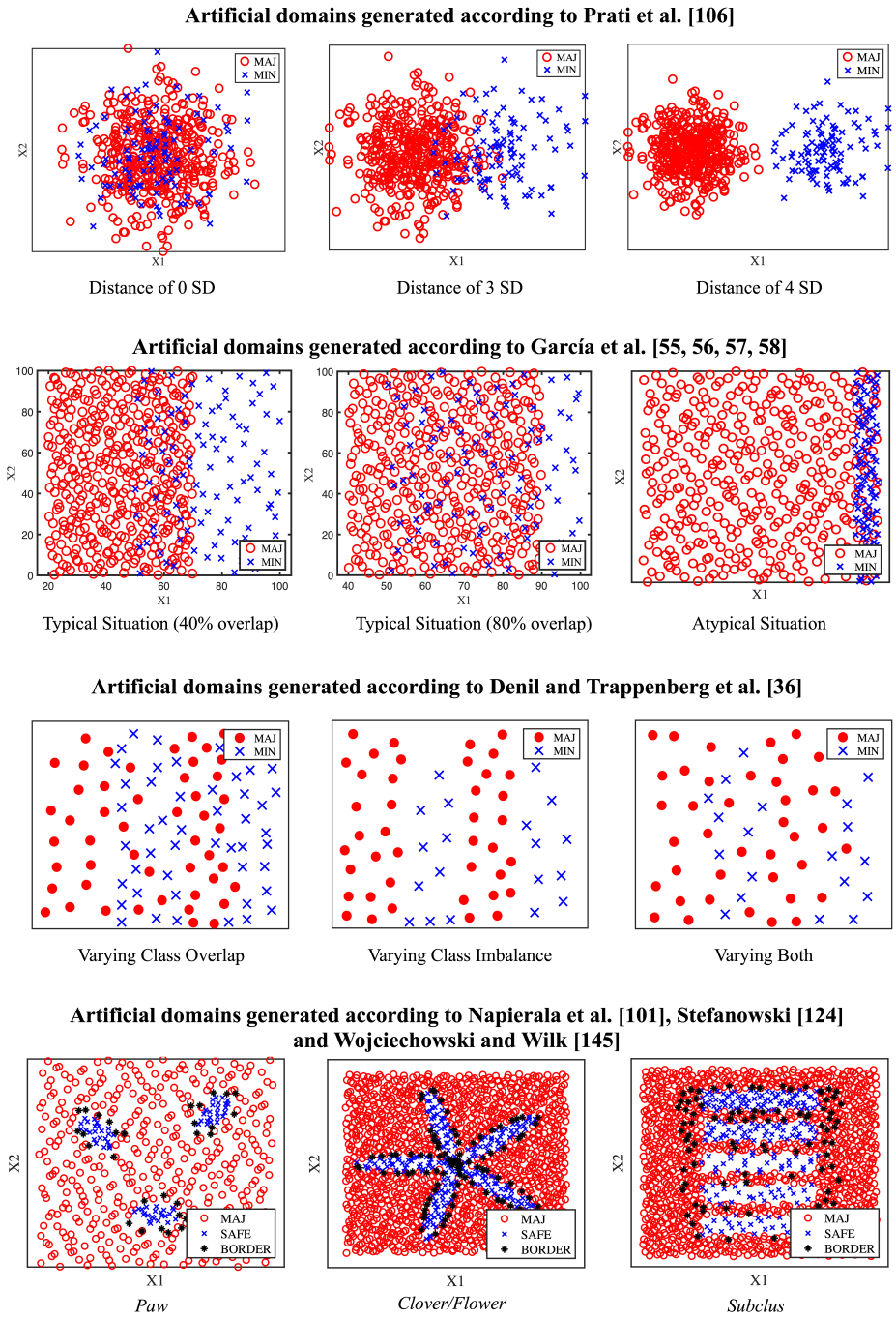

Fig. 1: Artificial domains considered in related work. The red circles represent the majority class examples (MAJ) while the blue crosses represent the minority class examples (MIN). Prati et al. [106] defined class overlap as the distance between cluster centroids of different classes. García et al. [55, 56, 57, 58] considered both typical and atypical configurations where class examples were distributed over squares of the same size. In typical domains, class overlap may either be determined as a fraction of the area that is overlapped over the total minority area, or over the total majority area. For atypical domains, class overlap was not quantified numerically. Denil and Trappenberg [36] divided the domains into four equal regions with alternating class memberships. Class overlap was captured by the extent to which adjacent regions intertwined. Napierala et al. [101], Stefanowski [124, 125], and Wojciechowski and Wilk [145] defined *paw*, *clover/flower* and *subclus* domains with increasing amounts of borderline minority examples (BORDER), represented by the black stars. Mercier et al. [98] reproduced several artificial data domains considered in previous works.

Table 1: Summary of existing literature on the joint-effect of class imbalance and overlap. For each related work are identified the objectives of the study, the used domains and controlled parameters, used classifiers, and major conclusions.

Study	Objective	Domains[2]	Controlled Parameters	Classifiers	Major Conclusions
Prati et al. 2004 [106]	Study the combined effects of class imbalance and overlap.	Two artificial Gaussian clusters (majority and minority) with unitary standard deviation (10.000 examples, 5 dimensions).	Distance between cluster centroids (1 to 9 standard deviations). Percentage of minority class examples (1% to 50%).	C4.5	Imbalance ratio is not the sole factor that affects classifiers: increasing amounts of class overlap significantly hinder performance results.
García et al. 2006 [55]	Study the effects of class imbalance and overlap on instance-based classification.	Two uniform squares of size 50 × 100 (majority and minority) with an IR of 4:1 (500 examples, 2D). Phoneme, Satimage, Glass and Vehicle datasets from UCI Repository.	Distance between square centres (6 different configurations). IR fixed at 4:1 and 500 examples.	1NN	The combination of imbalance and overlap causes a deterioration of classification performance.
García et al. 2007 [56]	Determine whether performance measures are able to distinguish between typical and atypical situations.	2-dimensional uniform squares creating typical and atypical situations.	Distance between square centres (Typical Situation). Density of examples (Atypical Situation). IR fixed at 4:1 and 500 examples.	1NN, MLP, NB, RBF, C4.5	Specificity and Sensitivity results seem to be good descriptors of the data complexity.
García et al. 2007 [57]	Study the behaviour of several classifiers on imbalanced and overlapped domains.	2-dimensional uniform squares creating typical and atypical situations.	Distance between square centres (Typical Situation). Density of examples (Atypical Situation). IR fixed at 4:1 and 500 examples.	1NN, MLP, NB, RBF, C4.5, SVM	Performance of classifiers is influenced by the type of situation (typical versus atypical).
García et al. 2008 [58]	Study the behaviour of KNN versus other classifiers, in typical and atypical situations.	2-dimensional uniform squares creating typical and atypical situations.	Distance between square centres (Typical Situation, IR 4:1). Density of examples (Atypical Situation, IR 4:1 and 50:1). 500 examples.	KNN, MLP, NB, RBF, C4.5	The class more represented in the overlap region is more easily recognised by global learning classifiers, while the class less represented in that same region benefits the most from more local classifiers.
Denil and Trappenberg 2010 [36]	Study the effects of class imbalance and overlap individually and in combination, with varying training set sizes.	Points generated in 4 regions with alternating class memberships, inside a square of length 1.	Overlap between classes (μ). Imbalance between classes (α). Size of training sets.	SVM	The combination of imbalance and overlap is more severe for classification performance than each factor taken individually. However, class overlap seems more prejudicial for classification than class imbalance. Increasing the training set size improves the classification performance when class imbalance is evaluated in isolation, yet degrading such performance when there is also class overlap.
Napierala et al. 2010 [101]	Study the impact of disturbing the borders of subregions of the minority class.	2-dimensional domains (*paw*, *clover/flower* and *subclus*) with 800 examples.	Percentage of borderline minority examples (0, 30, 50 and 70%). IR 7:1 and 800 examples.	C4.5, MODLEM	Increasing the percentage of borderline examples strongly deteriorates the performance of classifiers.

To be continued on the next page...

Table 1: Continued from previous page.

Study	Objective	Domains[2]	Controlled Parameters	Classifiers	Major Conclusions
Stefanowski 2013 [124]	Study the influence of overlapping in the boundary between classes (overlap was expressed as a percentage of borderline examples in the minority class).	2-dimensional domain (subclass data) with 800 examples.	Percentage of borderline minority examples (0, 10, 20%). IR 5:1 and 9:1 and 800 examples.	C4.5, Jrip, KNN	Besides the decomposition of the minority class, overlap is a critical factor that affects classification. Presence of class decomposition and overlap causes a larger performance deterioration than class imbalance.
Wojciechowski and Wilk 2017 [145]	Analyse the impact of class imbalance, data typology, and dimensionality in classification performance.	Artificial domains with varying shapes (*paw*, *clover/flower*) and dimensionality (2, 3, 5 and 7 dimensions).	IR (5:1, 7:1, 9:1, 13:1). Number of examples fixed to 1200 for *paw* and 1500 for *clover/flower*. Number of minority borderline examples (0% and 30%).	KNN, C4.5, PART, NB, RBF, SVM	Data typology is more critical than class imbalance and data dimensionality. KNN and SVM-RBF outperformed the remaining classifiers.
Mercier et al. 2018 [98]	Analyse the performance degradation of several classifiers in overlapped and imbalanced domains.	Artificial domains with varying shapes (*clusters*, *garcia*, *clover/flower*, *paw*, *subclus*) and dimensionality (2 to 40 dimensions).	IR (1:1, 2:1, 4:1, 6:1, 8:1 and 10:1). Percentage of minority safe and borderline examples (100/0 to 0/100) for *clover/flower*, *paw*, *subclus*. Distance between cluster centroids (*clusters*) and square centres (*garcia*). 1500 examples.	CART, KNN, FLD, NB, MLP, SVM	MLP and CART seem more robust to class overlap. KNN and linear SVM are the most aligned with *degOver*. Data dimensionality and structure/shape play an important role in explaining performance results.

[2] Some of the artificial domains discussed in related work are available at `http://keel.es`

3 Lessons learned on the characteristics of the data domains

From the analysis of related research, three main factors seem influential in synergy with class imbalance and overlap: local data characteristics, data structure and data dimensionality. We tackle each component independently to provide a summary of the most relevant findings and stress their significance.

3.1 Local Data Characteristics: Local Imbalance and Data Typology

In related work, the combination of class imbalance and overlap has different effects on the performance of classifiers, depending on the characteristics of the overlap region. In particular, the local imbalance in the overlap region is one of the most impactful factors [56, 57, 58]:

- When the class imbalance in the overlap region is the same as the global imbalance, classifiers with a more global nature tend to misclassify the minority examples as classes overlap, thus prioritising the majority class. Conversely, classifiers with a local nature make a decision regarding the class of examples based on their local neighbourhood, thus avoiding the bias towards majority concepts;
- When the minority class is dominant in the region of overlap, classifiers based on a more global learning obtain better results on the minority examples while more local classifiers work better for the majority class.

In sum, more global classifiers are able to better recognise the class more represented in the overlap region, whereas local classifiers perform better on the less represented class [58]. Note, however, that the dominance of a given class in the region of overlap illustrates a type of distribution skew [34]. In these situations, the results can be quite different from what is expected in standard imbalanced domains, such as the minority class obtaining better performance than the majority class (in the case of binary-classification problems), if the minority class is more represented in the overlap region. In the scenarios discussed in related work (atypical situations), the distribution skew is due to the local imbalance in the overlap region. However, distribution skews may arise irrespective of the class imbalance in the domain, e.g., they can be due to the data distribution/sparsity in the overlap region. They are, however, intrinsically related to the overlap between classes, and may give rise to particular representations of the problem, where the local characterisation of data is fundamental to fully understand the type of degradation created.

Data typology is also identified as one of the most important factors affecting classification performance in imbalanced and overlapped domains. The term "data typology" corresponds to a neighbourhood-based categorisation of examples into different types. Currently, four main categories are established and followed in recent works: *safe*, *borderline*, *rare*, and *outlier* examples [100]. It should be noted that although related work emphasises the number of minority borderline examples as relating to the problem of class overlap, other types of examples can also contribute to the whole overlap (e.g., non-safe examples, such as rare examples or outliers). With respect to data typology, the following insights may be derived:

- Data typology assumes a more influential role on the difficulty of classification tasks than class imbalance or data dimensionality [145];
- Increasing the number of borderline minority examples has shown to severely jeopardise the classification performance [124, 145], especially exacerbating the deterioration of tree and rule-based classifiers [101].

Overall, related research has systematically demonstrated that it is important to take the internal characteristics of the domains into consideration when studying the joint-effect of class imbalance and overlap. Herein, we highlight the importance of the local data characteristics in what concerns the existence of class distribution skews and different types of examples comprised in data. In fact, we acknowledge them as vortices of class overlap, i.e., existing representations of class overlap, as will be further discussed in Section 6.

3.2 Data Structure: Non-linear Class Boundaries and Class Decomposition

Let us first define the overall understanding of "data structure" taken in this paper. We treat the concepts of data structure, data shape and data morphology interchangeably. With these terms we refer to the structural properties of the data that comprise their form, the complexity of decision boundaries, and existing class decomposition. As an example, artificial domains in related work such as clusters [106], squares [58], *paw*, *clover/flower*, and *subclus* all possess different data structures, i.e., different morphologies, shapes, class decomposition, and class boundaries of different difficulty (Figure 1). To this regard, the following observations should be highlighted:

- More complex shapes are harder to learn, independently of the class imbalance and overlap characteristics. Under the same configuration of class overlap and imbalance, the classification performance has shown to be affected by the characteristics of the decision boundaries (e.g., squares versus concentric circles [58]);
- Domains presenting a complicated class decomposition are more difficult to handle: *subclus* domains are generally easier to learn than *paw*, which in turn are easier to learn than *clover/flower* domains;
- Tree and rule-based classifiers are especially affected by non-linear decision boundaries, whereas classifiers with other learning paradigms (KNN and SVM with a RBF kernel) do not seem as critically affected. Linear classifiers (FLD and SVM with linear kernel) are strongly affected by the data structure, with FLD often misclassifying all minority examples, irrespective of other factors (class imbalance, class decomposition, and dimensionality);
- The combination of complicated class decomposition and class overlap is more impactful for classification performance than the class imbalance for tree and rule-based classifiers, and KNN [124]. However, the effect of class overlap seems stronger than increasing class decomposition. This effect is especially critical for smaller datasets or non-linear class boundaries [124].

Complex data structures pose difficult challenges for classifiers, irrespective of other factors such as class overlap and imbalance. However, when occurring together with class overlap and imbalance, data structure acts as an exacerbator of

a complex problem in itself, amplifying the deterioration of classification performance. Non-linear decision boundaries require classifiers with a more local-based learning or kernel adaptations. In turn, class decomposition further relates to the problem of small disjuncts and the ability of classifiers to derive general or specialised rules [69]. It is therefore important to take these internal data characteristics into consideration when defining appropriate solutions for the identification and quantification of class overlap. This is especially true for real-world imbalanced domains, where the underlying class distributions and the number and structure of class concepts are unknown and difficult to discover or approximate.

3.3 Data Dimensionality

Although some research has focused on developing appropriate methods for dimensionality reduction in imbalanced domains [45], the combination of data dimensionality with other data characteristics has received very little attention in the literature. With respect to class overlap, since the majority of related work focuses on 2-dimensional domains, conclusions regarding data dimensionality are based on the research of Wojciechowski and Wilk [145], and Mercier et al. [98]:

- Overall, performance results improve with higher dimensionality. Additionally, increasing the class imbalance and class overlap seems to have a limited impact on the classification results;
- For domains with more complex data typology (i.e., not just increasing borderline examples but also rare and outlier examples), increasing the data dimensionality benefited the recognition of the minority class [145].

Class overlap seems to disappear as the dimensionality grows, which to some extent is related to changes in the data density for higher domains. If the total number of data examples is fixed, there will be a decrease of the data density as the dimensionality increases. For the domains studied in [98, 145] (*subclus*, *paw* and *clover/flower* domains), the majority class is especially affected, as it becomes sparser very rapidly. Consider for instance the *paw* domains, depicted in Figure 1. There are 3 well-defined minority class clusters (ellipsis) surrounded by an integumental space of the majority examples scattered across the remaining space. For higher dimensions, the minority clusters turn into hyper-ellipsis that become denser in comparison to the volume of the majority hyper-rectangle, thus improving class separability [145].

To this point, there is not much research on the evaluation of data dimensionality on imbalanced and overlapped domains. For instance, it remains unclear what would be the effects of dimensionality reduction techniques on the neighbourhood of data examples and consequently on their data typology and classification performance. For domains simultaneously affected by class overlap and imbalanced, feature selection is also an understudied topic, although some research has begun to shed some light on the subject [9, 30, 107]. These topics currently constitute open challenges for research.

4 Lessons learned on the nature of classifiers

Throughout related research, few works analyse the behaviour of classifiers beyond a comparison of classification performance results:

- In [58], authors distinguish between local (KNN) and global classifiers (MLP, NB, RBF, C4.5) and conclude that the performance of classifiers is related with the local imbalance of data in the overlap region, showing that a more local behaviour benefits the underrepresented concepts. Such behaviour is usually portrayed by instance-based classifiers, such as 1NN;
- In [145], classifiers are divided into symbolic (C4.5 and PART) and non-symbolic (KNN, NB, RBF, SVM). Symbolic classifiers lagged behind non-symbolic classifiers, although this may be due to the more extensive parametrisation of some non-symbolic classifiers (KNN and SVM performing the best);
- In [98], the performance degradation is associated with the learning paradigm of each classifier. Classifiers that work on the basis of data space fragmentation (CART, MLP, and KNN) seem less affected by class overlap, whereas linear classifiers (FLD and SVM-linear) perform the worst.

Understanding how the joint-effect of class overlap and imbalance (as well as data characteristics) affects the performance of each classifier is a step towards the definition of adequate strategies to handle the problems simultaneously. Overall, related work has shown that major differences between the performance of classifiers rely on their ability to provide specialised decisions, where local learning paradigms have shown to be better suited to several sources of complexity, such as distributions skews, difficult data typologies, and complex data structures:

- Among all families of classifiers, instance-based classifiers (KNN) have shown to be the most resilient to changes in class imbalance and overlap. Throughout related research, KNN was able to achieve good results even for difficult situations characterised by class distributions skews [58], and complex data typology [145]. Its sensitivity to changes in local imbalance, and flexibility for complex data structures, turn it into a simple, yet efficient, approach to study the combination of class imbalance and overlap;
- Other classifiers have also shown to be adequate choices to handle issues simultaneously. RBF networks and SVM with RBF kernel have shown to be robust to distributions skews and difficult data types, as well as more complex shapes. Conversely, NB, although showing a high tolerance to class overlap and performing successfully in distribution skews and complex domains, is somewhat affected by class imbalance and difficult data types [58, 98, 145];
- Linear classifiers, and rule and tree-based classifiers obtained lower performance results, presenting some limitations under several sources of complexity.

In what follows, we will focus on distinct families of classifiers and their learning paradigms, aiming to provide an overview of their behaviour under imbalanced and overlapped domains. In that sense, we consider four main families: *Instance-Based Classifiers*, *Rule and Tree-Based Classifiers*, *Bayesian Classifiers*, *Neural*

Networks, and *Support Vector Machines and Linear Discriminants*. For each family of classifiers we highlight the most important findings from related work[3].

Instance-Based Classifiers (KNN)

- As KNN presents a local nature, it effectively addresses regions with different local data densities, i.e., it does not present the general bias towards the most represented class as most global classifiers;
- Smaller values of k guarantee its local nature and allow a more successful recognition of less represented concepts in the overlap region. In turn, for larger values of k, KNN approaches the behaviour of more global classifiers, which benefits the more represented concepts in that region [58];
- Considering higher values k has also proven beneficial for the recognition of the minority class when the number of borderline minority points in the overlap region increases [145];
- The local behaviour of KNN is also advantageous for more complex data structures (non-linear shapes), where KNN is among the top performers, irrespective of the class imbalance and class decomposition [98, 124, 145].

Rule and Tree Classifiers (C4.5, CART, PART, MODLEM)

- Class overlap highly degrades the performance of rule and tree-based classifiers, more than class decomposition [101, 124]. Additionally, a faster performance deterioration is observed for more complex non-linear shapes;
- MODLEM outperforms C4.5 when compared under the same conditions (borderline minority examples, class decomposition and imbalance ratio) [101]. Also, CART outperforms C4.5 even for higher percentages of minority borderline examples and imbalance ratios [98, 101]. We hypothesise that this difference may be due to the splitting criteria;
- Both pruned and unpruned versions of C4.5 and PART obtain nearly the same results for the same amount of class overlap (borderline examples), although for more difficult types of examples (rare and outlier examples), unpruned versions generally perform better [145].

Bayesian Classifiers (NB)

- NB has performs successfully for both typical and atypical domains [58] and more complex data shapes [98, 145];
- In [145], although NB is successful in classifying datasets with increasing amounts of borderline minority examples, it performs poorly for more difficult types (rare and outlier examples);

Neural Networks (RBF, MLP)

- For typical and atypical domains [58], RBF and MLP obtain similar results. However, for more complex shapes (atypical concentric circles), MLP fails to recognise all minority examples whereas RBF network provides similar results to atypical domains. This difference may reside on the activation function of each network. MLP uses a sigmoid activation function, whereas RBF uses a Gaussian activation function, which makes neurons more locally sensitive [68].

[3] The interested reader may find detailed information on the performance of each classifier in the supplementary material provided online at `https://student.dei.uc.pt/~miriams/pdf-files/AIR_2021_Appendix.pdf`

- RBF also shows a good performance for *paw* and *clover/flower* domains, being among the top performers [145]. MLP handles *clover/flower* domains better than *subclus* domains, although the former shape is considered more complex [98]. We hypothesise that this could be due to the fact that *clover/flower* is a unified shape: the subregions are connected and have similar densities. In turn, *subclus* has 5 disconnected subregions with different densities. For MLP, learning five decision boundaries with different densities seems more difficult than to learn a single (although complex) decision boundary with an even representation of points among subregions. For *subclus* domains, class overlap seems to affect MLP classification performance more than class imbalance, whereas for *clover/flower*, class imbalance seems the most prejudicial [98];

Support Vector Machines and Linear Discriminants (SVM and FLD)

- SVM is more deeply affected by class overlap than class imbalance, although the combination of both problems is even more costly [36]. SVM further exhibits a breaking point occurring when nearly half of the domain is overlapped and the imbalance ratio in the overlap region approaches a balanced scenario [36, 58];
- In [98, 145], SVM shows a competitive performance for increasing amounts of borderline minority examples. The good behaviour of SVM is associated with the tuning of hyperparameters performed.
- Both linear SVM and FLD are extremely affected by the structure of data. In particular, FLD fails to classify any minority examples for domains with non-linear decision boundaries, although it performs reasonably well for more simple shapes (typical square domains or cluster domains) [98]. FLD aims to find a projection onto a line (one-dimensional space) where classes are well separated, which for non-linear class boundaries is extremely difficult;
- Contrarily to the remaining classifiers, the increase of data dimensionality does not seem to improve FLD in the classification of non-linear decision boundaries. Although the generation of overlap in higher dimensions increases concept separability, the projections performed by FLD remain compromised.

With respect to the top performing classifiers, note how the hyperparametrisation plays a vital role, especially with the use of Gaussian kernels. Although KNN, SVM-RBF and RBF networks are based on different learning paradigms, by using Gaussian kernels, SVM-RBF and RBF can approximate the local behaviour of KNN, depending on the chosen hyperparameters. Hyperparametrisation can help solving issues simultaneously by defining appropriate parameters depending on the characteristics of data. As an example, different parametrisations of KNN could be used to solve successfully domains with distribution skews for all classes, by choosing smaller values of k in regions where a given class is sparse or less represented and larger values when a class is dense or well-represented in overlapping regions. The same can be derived for kernel parameters.

This remains an understudied topic in imbalanced and overlapped domains and is currently an open direction for future research. The main idea is that attending to the bias of classifiers and the representation of class overlap in the domain, one can establish appropriate strategies to improve classifiers individually (as is the

case of improving parametrisation for different regions) or combining local and global classifiers to achieve improved performance (e.g., via ensemble learning, where the choice of individual classifiers may be tailored to the characteristics of the data domains). Naturally, this requires a full characterisation of the overlap problem in imbalanced domains, which to this point is not a well-established topic in the literature, as we will further detail in the following section.

5 Limitations of Seminal Research

Despite related research provides interesting findings, as discussed throughout the previous sections, there is still a long way to go before extrapolating insights for real-world domains. Indeed, related research has the following limitations:

- All research works consider artificially generated data domains, where class overlap, class imbalance, data typology, class decomposition, local data densities, and data dimensionality are defined *apriori*;
- Not all aspects are studied across all research works: class decomposition and data dimensionality are understudied. Also, authors often neglect scenarios of extreme imbalance;
- Experiments are confined to well defined shapes (e.g., squares or clusters of data), with little minority class decomposition (maximum of 5 subregions for *clover*/*flower* and *subclus* domains), a regular majority class representation (an integumental region, without class decomposition) and small data dimensionality (most works are limited to 2-dimensional domains);

Naturally, control over these parameters allows a better understanding of the generated domains and consequently a more precise evaluation of obtained results. Also, the insights provided over synthetic data lay the foundation for the interpretation of results over real-world domains, and respective investigation of specialised approaches. This was the rationale behind the thorough analysis of previous research that culminated in the insights summarised in Sections 3 and 4. To this regard, the conclusions derived previously are to be taken as a global view on the peculiarities of the data domains and footprints of classifiers, showing that the combination of class imbalance and overlap may give rise to a multitude of scenarios, each presenting its own implications for classification tasks in general, and classification paradigms in particular. Nevertheless, generalisation for real-world datasets requires further investigation and it is important to discuss some open issues that prevent that more profound conclusions are derived:

Class overlap is not mathematically well-established:
Throughout related research, there is no standard measurement of the overlap degree. Hence, class overlap is measured in rather distinct ways. Prati et al. [106] measure class overlap as the distance between cluster centroids, which does not reveal the exact degree of overlap in each configuration. Similarly, the research of García et al. [55, 56, 57, 58] lacks a formulation of the overlap degree. Given the simplicity of typical domains, one may infer that the degree of overlap can either be determined as a fraction of the area that is overlapped over the total minority area, or over the total majority area. However, for atypical situations, the notion of overlap degree gets rather lost (no percentages or any other values are presented for the overlap degree) and the results

need to be evaluated considering the local imbalance combined with the size of the overlap region, instead of evaluating an exact measure of class overlap. Furthermore, these methods of estimating class overlap do not generalise for different data structures (e.g., non-geometrical shapes) or for a higher number of dimensions, frequently found in real-world domains. Although it may seem an intuitive concept, to this point there is not a well-established mathematical definition for class overlap [135]. This may be due to the fact that, as the literature progresses, several concepts associated with class overlap have been brought to light, leading to the discussion of distinct representations of the problem.

Class overlap assumes different representations:
In related work, class overlap is often associated to different concepts, that ultimately result in its characterisation according to different representations. Class overlap is often associated to concepts such as class separability (distance between cluster centroids [106]), overlapping regions or areas [55, 56, 57, 36], structural biases such as distribution skews (local imbalance in overlapping regions) [58], complex structures (class decomposition, data sparsity [101, 124, 145]), data typology (via borderline examples [101]), and the discriminative power of features (data dimensionality [98, 145]). These representations of class overlap are assessed differently (e.g., distance between concepts, percentage of overlapped area, combination of local imbalance with size of overlap region, percentage of borderline examples), which complicates the comparison of results among related work. Also, except for data typology, the used measures for the assessment of other overlap representations are not generalisable for real-world domains. Identifying and quantifying class overlap becomes a more strenuous task if it has different representations. Different representations of class overlap are associated with different insights regarding the domain and represent different sources of degradation. However, to this point, no study in the literature refers to this issue. What is more, studying class overlap without measuring it clearly (not to mention without attending to its different representations) may prevent meaningful insights from being derived: general conclusions can be obtained (i.e., with respect to the overall effect of class overlap), but it is not possible to extract more specific guidelines for future developments in the field.

The class overlap degree does not take other factors into account:
Prati et al. [106] control class overlap as a distance between clusters centroids, although it does not take into account the data sparsity in the overlap region, which conditions the number of examples that effectively contribute to class overlap. Similarly, when García et al. [58] measure class overlap as a percentage of overlapped area, the distribution of examples within the overlapping area is not considered. For instance, two typical domains with different global class imbalance may have the same overlap area, although the number of data examples in the overlap region is different. If we were to consider atypical domains, the issue is even more clear. Note how both a typical and atypical situation may have the same overlap area, although they refer to two very distinct situations in terms of class overlap and associated difficulty for classification tasks. Furthermore, recall that in related work, atypical situations do not have an associated measure. As discussed in Section 3, the local properties

of data are important to characterise the degradation that the class overlap produces. To this regard, situations presenting class skews (generated by data distribution/sparsity, or local imbalance) are important to acknowledge when producing an overlap measure. Napierala et al. [101], Stefanowski [124, 125], and Wojciechowski and Wilk [145] consider the local characteristics of data by associating class overlap to the percentage of borderline minority examples in the domain. Nevertheless, depending on how they are distributed, two domains with the same percentage of borderline minority examples may affect the classification tasks differently. In addition, despite borderline examples are highly related to the problem of class overlap (closer to class boundaries), other examples scattered throughout the domain may also contribute to class overlap.

Due to these limitations, we argue that the joint-effect of class overlap and imbalance is still not fully characterised. One may argue that, since seminal work on this topic, other approaches have been attempted to define a more accurate characterisation of domains and its relation with classification performance. A natural question therefore arises: "Moving past seminal work, how is the combination of class imbalance and overlap currently handled in real-world domains?" To shed some light on this matter, the following sections elaborate on two important aspects. One is the identification and quantification of class imbalance and overlap, whereas the other is the devise of suitable techniques to overcome these issues simultaneously (both focusing on real-world domains). We therefore provide a comprehensive analysis of measures to characterise class imbalance and class overlap (Section 6), and a thorough overview of the state-of-the-art class overlap-based approaches used in imbalanced domains (Section 7). We will show that, despite the recent developments in the field, the measures and approaches devised for real-world domains still suffer from similar limitations as previous research on synthetic data. This will be made clear throughout the following sections, motivating our claim regarding the need to move towards a unified view of the class overlap problem in imbalanced domains.

6 A Taxonomy of Class Overlap Measures

Throughout the years, class imbalance has been consistently estimated by considering the number of examples of each class and computing the Imbalance Ratio (IR), such as $IR = 2$ or $IR = 2:1$ (Equation 1), or determining the percentage of minority class examples in the domain (Equation 2) (note that we are focusing on binary-classification problems for simplicity, extensions for multi-class domains can be found in [88]). Other definitions of class imbalance can be found in [87] (Entropy of Class Proportions), [120] (Minority Value and Class Balance), and [98] ($degIR$). These measures are, however, only discussed within the respective papers, whereas IR and Minority (%) represent the formal, well-established definitions accepted in the field [46].

$$\text{IR} = \frac{|C_{maj}|}{|C_{min}|} \tag{1}$$

, where $|C_{maj}|$ and $|C_{maj}|$ represent the number of majority and minority examples in the domain, respectively.

$$\text{Minority } (\%) = \frac{|C_{min}|}{N} \times 100 \tag{2}$$

, where N represents the total number of examples in data.

On the contrary, estimating class overlap is a more complicated task, given that it comprises several representations, as discussed in Section 3. Indeed, certain intrinsic characteristics of data (class imbalance, local imbalance, data typology, non-linear boundaries, class decomposition, data dimensionality) may give rise to different facets and degrees of overlap. Before focusing on specific measures and approaches, let us discuss some situations to clarify the idea that class overlap may comprise different representations and that the overlap degree may be affected by other factors, namely class imbalance. Herein we will briefly refer to some measures of class overlap to discuss this issue, but they will be thoroughly described in the following sections.

We start by analysing the synergetic effects of class imbalance and overlap over the domains presented in Figure 2, previously discussed in seminal work [58] (Section 2). Figure 2 represents two "typical situations", where classes are uniformly distributed over 2-dimensional squares with the same size. In these domains, the computation of the class overlap degree was either determined as a fraction of the area that is overlapped ($A_{overlap}$) over the total minority area (A_{min}), or over the total majority area (A_{maj}), since $A_{min} = A_{maj}$. As an example, consider the scenario depicted in Figure 2 (left-side), where the domain presented a class overlap of 40% [58]. This overlap percentage may be calculated as $\frac{A_{overlap}}{A_{min}} \times 100$ or $\frac{A_{overlap}}{A_{maj}} \times 100$, which corresponds to an overlap degree of $\frac{2000}{5000} \times 100 = 40\%$.

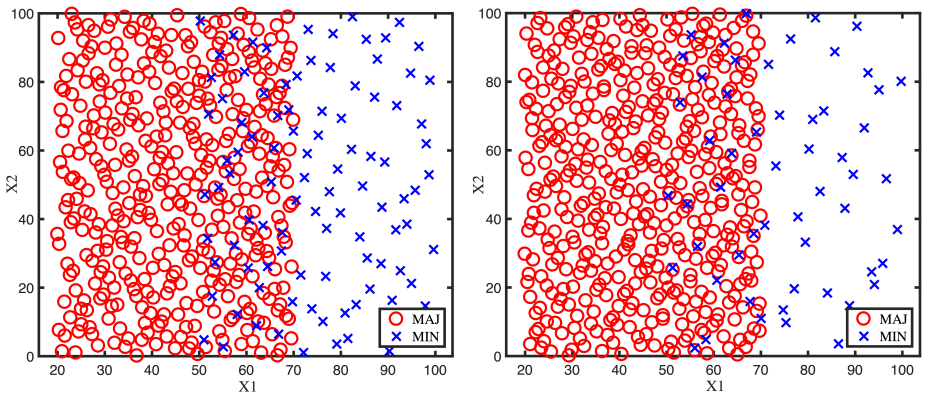

Fig. 2: Artificial domains generated according to García et al. [55]. Although the overlap region is the same in both examples, one domain (left-side) considers an IR of 4:1 whereas the other (right-side) has an IR of 8:1. According to the percentage of overlapped area, both reveal the same overlap degree (40%), although due to the imbalance ratio, the local properties of domains are rather different.

Now, note how focusing a measure of class overlap solely on the area of the overlap regions does not take the imbalance ratio into account. For instance, in Figure 2 (left-side), the domain is generated for an IR of 4:1, for 500 examples: would it be adequate to assume that the same setup for a 8:1 ratio (Figure 2, right-side) would also produce a class overlap of 40%? Since the number of conflicting examples in the same overlap region is lower, this may not be the case. Nevertheless, measuring class overlap as a percentage of the overlapped area remains a common strategy used in the experimental setup of recent research [133, 135]. Note also that determining the number of misclassified examples following a k-Nearest Neighbour rule (another strategy to quantify class overlap, more closely related to the concept of local data characteristics - to be discussed in Section 6.3), would return a different overlap degree for each scenario, whereas determining the size of overlapped area is more related to the structural properties of the data, and unable to capture more local changes in the domain. The key idea here is to show how class overlap may depend on other characteristics (class imbalance in this example) and that different measures capture different representations/vortices of class overlap.

Let us consider another example on different facets of class overlap, by examining Figure 3. The example shows two scenarios where class overlap is measured according to the Maximum Fisher's Discriminant Ratio, F1 (discussed in Section 6.1). In both scenarios, the data is projected onto the axis of features f_1 and f_2. The projections are the same for the f_1 but differ for f_2. Since F1 is maximal (and the same) in both situations, the scenarios reveal the same class overlap degree. However, in the scenario to the right, the separability of f_2 increases when compared to the situation to the left. If local information is taken into account, this domain would return a different overlap degree, since the number of misclassified examples (1NN) is lower (misclassified examples are marked in grey in Figure 3). Additionally, F1 does not consider class imbalance: for two datasets with different imbalance ratios and similar statistical properties (i.e., means and variances of each class are similar for both scenarios), F1 returns similar values. Again, this shows that class overlap may comprise different representations and that certain measures are able to capture some while failing to uncover others. In this case, F1 focuses on feature-level overlap, but does not consider local data characteristics (local information).

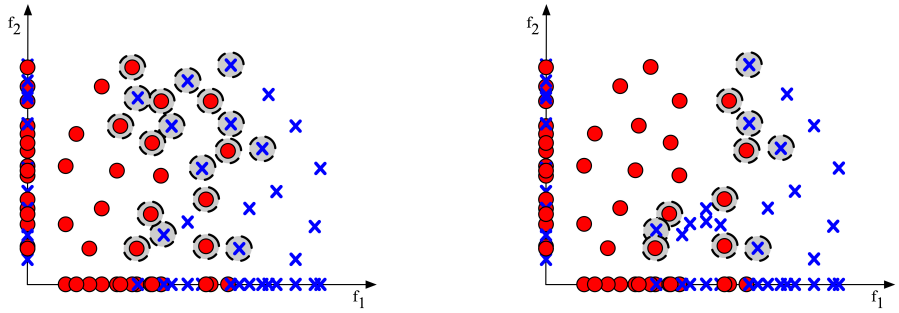

Fig. 3: F1 measures the highest discriminative power for all features in data, i.e., it returns the minimum overlap of individual features found in the domain. Accordingly, the scenarios above reveal the same discriminative power: feature f_1 has the same (and highest) F1 value in both cases. However, the individual overlap in feature f_2 is different, which makes these scenarios different in terms of classification difficulty. F1 therefore captures one facet of class overlap (feature overlap) but it does not provide a full characterisation of the class overlap problem in the domain.

Now that we have established that class overlap may comprise several representations and that some measures are able to capture some representations while neglecting others, it is important to establish the link between existing measures of class overlap in the literature, and the type of information (vortices of class overlap) they are associated to.

Throughout the years, several measures have been proposed and reformulated to identify and estimate certain properties of the data domains, referred to as *data complexity measures* [66, 88, 104, 121]. The most well-known taxonomy of complexity measures is the one defined by Ho and Basu [66], although throughout the years, other authors sought to complement this taxonomy, presenting their own division or proposing additional categories [88, 121]. Overall, these measures provide important insights regarding several properties of data and naturally, some relate to the problem of class overlap. However, complexity measures often focus on individual characteristics of the data, which might be insufficient to fully characterise class overlap, given that it is a heterogeneous concept comprising different sources of complexity (especially in the presence of other factors, such as class imbalance). A first step towards a robust characterisation of class overlap would be the definition of a taxonomy of class overlap measures that attends to its different representations, i.e., sources of complexity. However, although class overlap is considered one the most harmful issues for classification problems [43, 58], no such taxonomy currently exists. In what follows, we propose a novel taxonomy of complexity measures for class overlap, focusing on different vortices/representations of the problem and the measures that are able to characterise them.

Our taxonomy of class overlap complexity measures comprises four main groups: measures associated to Feature Overlap, Structural Overlap, Instance-Level Overlap and Multiresolution Overlap. Figure 4 provides an overview of the proposed taxonomy, where each group is established depending on the representation of class

overlap it is more suited to capture. Also, the concepts associated to each representation are highlighted and the measures for which adaptations to imbalance domains have been explored in the literature are identified. The following sections thoroughly characterise each group and their respective class overlap measures. All measures described in this section are implemented in a new Python library named `pycol` - *Python Class Overlap Library*, publicly available on GitHub[4].

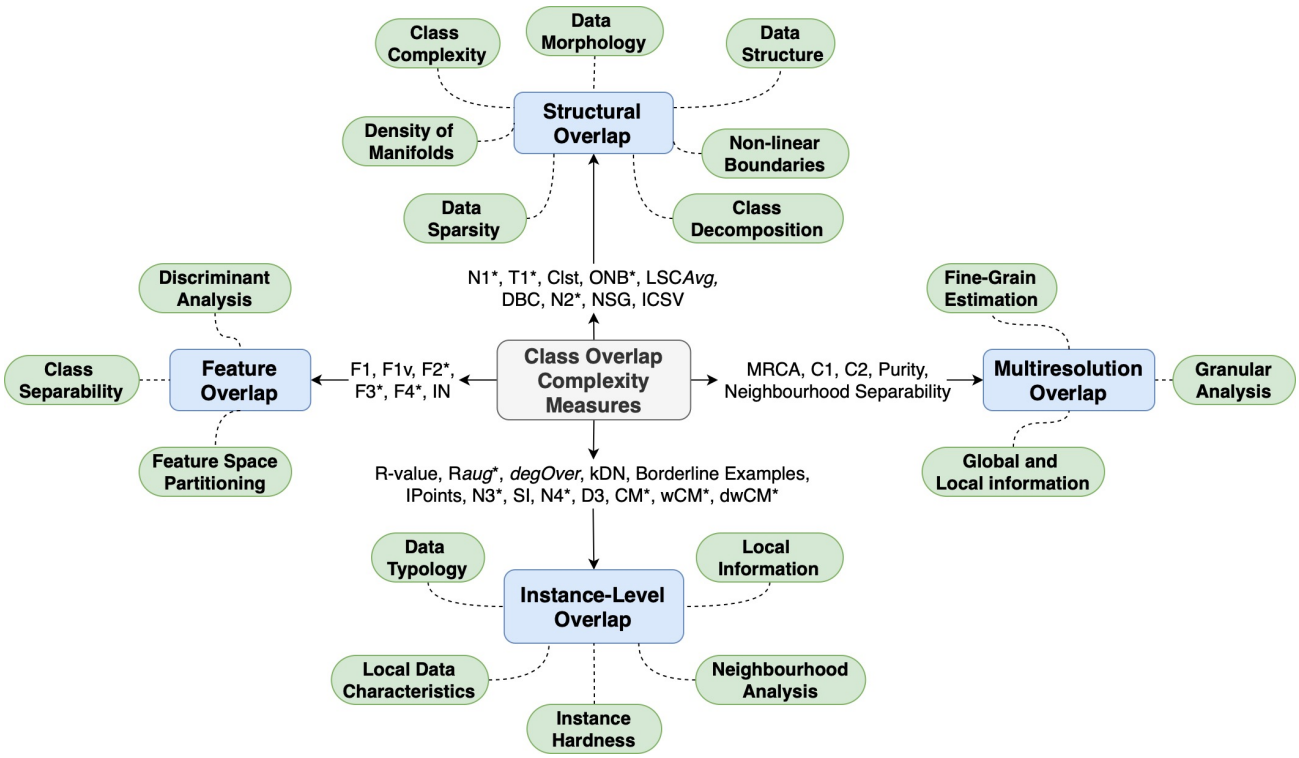

Fig. 4: Taxonomy of class overlap complexity measures. Different groups can be established depending on the representation of class overlap they are attentive to. Measures marked with an asterisk are those for which adaptations to imbalanced domains have been discussed in the literature.

6.1 Feature Overlap

These measures characterise the class overlap of individual features in data. Some are deeply associated to the concept of class separability, i.e., individual feature separability (F1, F1v) and focus on certain properties of class distributions to determine the discriminative power of features. Others recur to feature space par-

[4] `https://github.com/miriamspsantos/pycol`

titioning to delimit overlap regions (F2, F3, F4, IN), i.e., they divide features into certain ranges where data overlap is analysed.

6.1.1 Maximum Fisher's Discriminant Ratio (F1)

The maximum Fisher's discriminant ratio (F1) is perhaps the widest used measure to compute the overlap degree of a given dataset [86, 89, 111].

For each feature f_i comprised in the dataset, the Fisher's discriminant ratio (r_{f_i}) is obtained through Equation 3, where μ_1, μ_2, σ_1^2, and σ_2^2 are the means and variances of class 1 and 2, respectively. Then, F1 is obtained by finding the maximum r_{fi} over all features in data. As depicted in Figure 5 (to the left), F1 traditionally measures how discriminative each feature is, i.e., how well it can separate classes. Intuitively, higher values of F1 indicate less overlapped domains.

$$r_{f_i} = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{3}$$

In order to provide a measure of class overlap rather than class separability, Lorena et al.[88] establish the inverse of the original F1 formulation: $F1 = \frac{1}{1+r}$, where r is the maximum r_{fi} among all features. In such a case, higher values of F1 indicate more overlapped domains.

6.1.2 Directional Vector Maximum Fisher's Discriminant Ratio (F1v)

Rather than determining the separability of classes on the projection of data perpendicular to the axes (please refer to Figure 5), F1v searches for a vector where data can be projected with maximum separability [104]. It computes the two-class Fisher criterion, dF, as defined in Malina [92], where higher values indicate a higher separability between classes. Similarly to F1, Lorena et al. [88] define F1v as follows from Equation 4, where lower values indicate that there is a vector capable of separating classes after projecting data onto it. In other words, higher values of F1v indicate higher amounts of class overlap.

$$\text{F1v} = \frac{1}{1 + dF} \tag{4}$$

6.1.3 Volume of Overlapping Region (F2)

To determine F2, the overlap of the distribution of feature values is computed individually for each feature ($f_i = 1, \ldots, m$). First, the maximum and minimum values of each feature f_i are found, considering both classes C_1 and C_2. Then, the overlap length of feature values is determined and normalised by the overall range of the feature. Finally, F2 is determined by multiplying the ratio obtained for each feature (Equation 5), where higher values indicate a greater amount of class overlap. An example of the determination of F2 is depicted on Figure 5 (rightside).

$$\text{F2} = \prod_{i=1}^{m} \frac{overlap(f_i)}{range(f_i)} = \prod_{i=1}^{m} \frac{\max\{0,\, \text{minmax}(f_i) - \text{maxmin}(f_i)\}}{\text{maxmax}(f_i) - \text{minmin}(f_i)}, \text{ where} \tag{5}$$

$\text{minmax}(f_i) = MIN\big(max(f_i, c_1), max(f_i, c_2)\big)$,
$\text{maxmin}(f_i) = MAX\big(min(f_i, c_1), min(f_i, c_2)\big)$,
$\text{maxmax}(f_i) = MAX\big(max(f_i, c_1), max(f_i, c_2)\big)$,
$\text{minmin}(f_i) = MIN\big(min(f_i, c_1), min(f_i, c_2)\big)$.

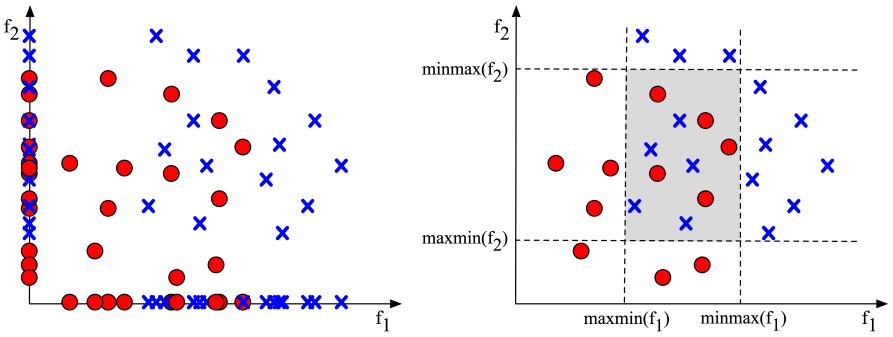

Fig. 5: Representations of F1 (leftside) and F2 (rightside) measures for the same dataset. Note how F1 projects data onto the axis to establish the amount of overlap, where f_1 is the feature with highest discriminative power, i.e., lowest overlap. In turn, F2 considers both features to define a region where classes coexist.

6.1.4 Maximum Individual Feature Efficiency (F3)

Traditionally, F3 measures the discriminative power of individual features by determining the efficiency of each feature and returning the maximum value [66]. For each feature, F3 determines the regions where there are values from both classes and then returns the ratio of feature values that are not in the overlapping regions. In Lorena et al. [88], a complementary measure is presented, where F3 measures the minimum amount of overlap between feature values of different classes (Equation 6). Thus, higher values of F3 indicate more overlapped domains (Figure 6).

$$\text{F3} = \min\Big(\frac{n_{overlap}(f_i)}{n}\Big) \tag{6}$$

, where $i = 1, \ldots, m$ features and n is the total number of examples in data.

$$n_{overlap}(f_i) = |\{x_j \in f_i : x_j > \text{maxmin}(f_i) \ \wedge \ x_j < \text{minmax}(f_i)\}| \tag{7}$$

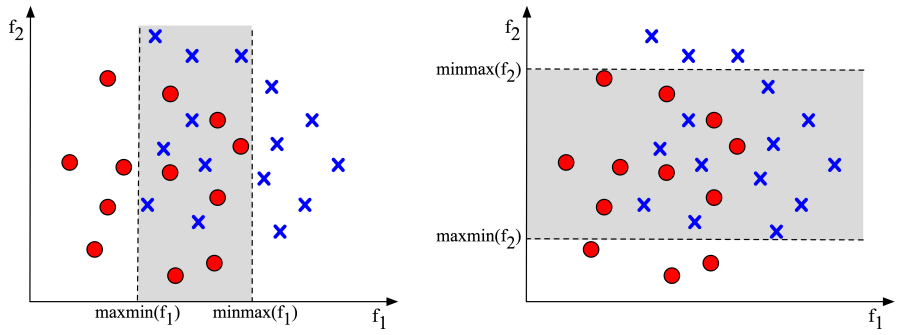

Fig. 6: Representation of F3 measure for the data domain of Figure 5. Feature efficiency is measured individually for f_1 (leftside) and f_2 (rightside), where f_1 is the most efficient feature, i.e., it returns the minimum amount of overlap. Adapted from [88].

6.1.5 Collective Feature Efficiency (F4)

Whereas F3 focuses on individual feature efficiency, F4 considers the discriminative power of all features [104]. To find F4, the following procedure is applied: first, the feature with highest discriminative power (lowest overlap) according to F3 is taken and all examples that can be separated using this feature are removed from the data. Then, the next most discriminative feature (considering the remaining examples) is taken and the process is repeated iteratively over all features. In the end, according to the original formulation [104], F4 returns the proportion of examples that have been discriminated, thus providing an idea on the proportion of examples that could be correctly separated by hyperplanes parallel to one of the axis of the feature space. Lorena et al. [88], however, consider F4 as the ratio of examples that could not have been separated (Figure 7). Thus, higher values of F4 indicate a larger amount of overlap between classes, considering all features collectively. F4 may be determined by Equation 8, where f_l represents the last most discriminative feature found through the iterative process described above and n is the total number of examples in data.

$$\text{F4} = \frac{n_{overlap}(f_l)}{n} \tag{8}$$

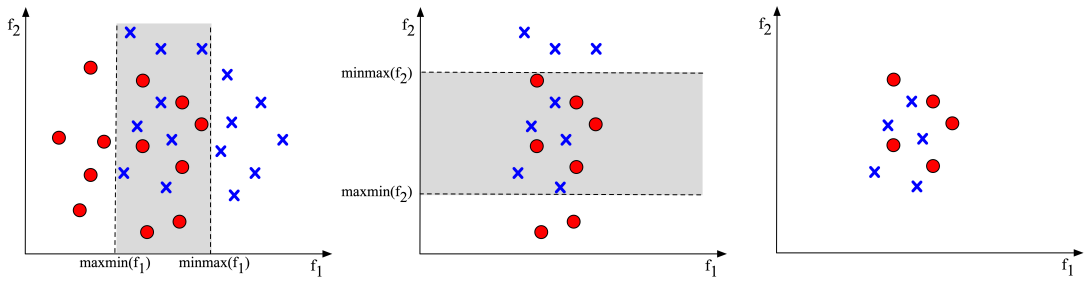

Fig. 7: Representation of F4 measure for the data domain of Figure 5. Since f_1 is the most efficient feature, all examples that can be separated according to f_1 (outside the grey area) are removed. Then, the same is performed on f_2. The remaining data examples are those who could not be separated, thus contributing to class overlap. Adapted from [88].

6.1.6 Input Noise (IN)

The Input Noise (IN) is related to the amount of overlap between features of different classes [136]. To determine the input noise, the maximum and minimum values of each feature for each class are used to define their boundaries. Then, if a given example falls inside the boundaries of another's class feature values, it is contributing to the overlap on this feature. To this regard, the input noise is related to F2 and F3 measures. However, the input noise measure then determines, for each example, in how many dimensions (features) it overlaps and normalises the total by $n \times D$, where n is the number of examples in data and D is the number of existing dimensions (Equation 9). Higher values of IN indicate higher amounts of class overlap. In Equation 9, g_i represents the number of features where the i^{th} example is in overlapping regions.

$$IN = \frac{1}{n \cdot D} \sum_{i=1}^{n} g_i \tag{9}$$

6.2 Structural Overlap

This group of measures is associated with the concept of class complexity (non-linear boundaries and class decomposition), comprising information on the internal structure of classes (data morphology). They can be used to characterise class overlap regions using a "divide-and-conquer" perspective, i.e., focusing on the structure of the domain to find problematic regions. Some measures analyse the properties of a Minimum Spanning Tree (MST) built over the data domain to produce measures of decision boundary complexity and structural overlap (N1). Others approach the identification of class overlap using the notion of hypersphere coverage (T1, $Clst$, ONB, LSCAvg). Some consider both MST and hypersphere coverage (DBC). Finally, also linked to the concept of data morphology, other measures aim to quantify the data sparsity/density of manifolds (N2, NSG, ICSV).

6.2.1 Fraction of Borderline Points (N1)

N1 measures the proportion of examples that are connected to the opposite class by an edge in a Minimum Spanning Tree (MST) [66]. Most often, these examples are those located near the boundary between classes, or those inserted in overlapped regions in the data space. In general, higher values of N1 indicate a higher degree of class overlap (classes are more deeply intertwined) [88, 105]. However, there are situations where N1 may assume higher values for simpler domains, e.g., if the class boundary has a narrower margin than the intra-class distances [111]. Considering V and E as the set of vertices and edges of a $MST(V, E)$, N1 can be defined by Equation 10, where y_i is the class label of a given example x_i.

$$\mathrm{N1} = \frac{1}{|V|}|\{x_i \in \mathrm{V} : \exists (x_i, x_j) \in \mathrm{E} \wedge y_i \neq y_j\}| \tag{10}$$

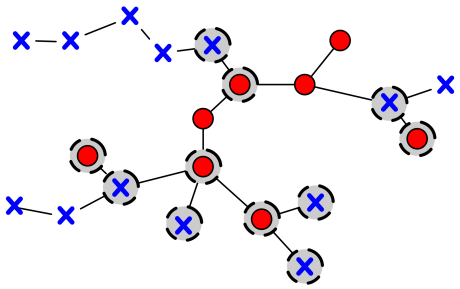

Fig. 8: A representation of the N1 measure. Marked points from both classes are those contributing to class overlap (connected to the opposite class in the MST). Adapted from [88].

6.2.2 Fraction of Hyperspheres Covering Data (T1)

To determine T1, a hypersphere centred at each example of the dataset is created and its radius is grown until it reaches an example of the opposite class. Then, hyperspheres contained in larger ones (of the same class) are eliminated (Figure 9). T1 is then defined as the ratio of hyperspheres that remain, as shown in Equation 11, where n represents the total number of examples in data.

$$\mathrm{T1} = \frac{\#Hyperspheres}{n} \tag{11}$$

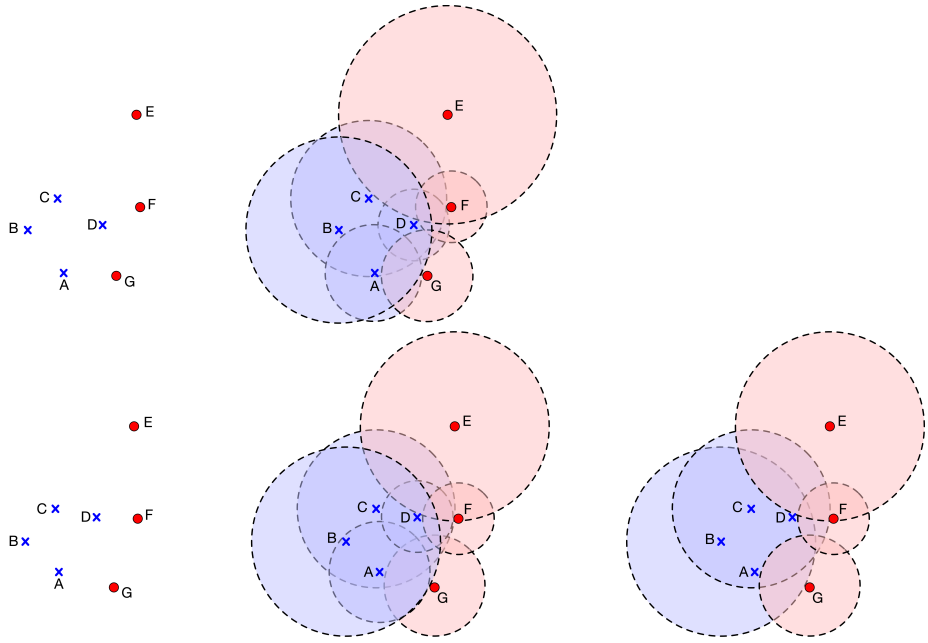

Fig. 9: Representation of the original T1 solution for two datasets (top and bottom rows). In the scenario depicted in the top row, the hyperspheres of points D and A are not completely absorbed by any other hypersphere in the domain. On the contrary, in the scenario of the bottom row, hypersphere D and A are absorbed by hyperspheres C and B, respectively, and are therefore eliminated.

Lorena et al. [88] consider an alternative implementation of T1, where the growth of a hypersphere is stopped when it starts to touch a hypersphere of the opposite class. Accordingly, this modification starts by determining the existing mutual nearest enemies in data, for which their radii are automatically established as half of the distance between them. The radius of the remaining hyperspheres are then determined recursively (Figure 10).

Given that the hyperspheres only contain examples of the same class, higher values of T1 indicate a larger amount of class overlap. Nevertheless, this measure is also sensitive to the distribution of data in the domain, i.e., covering situations where the domain is composed by different clusters of the majority and minority classes (even if there is no class overlap), will require a higher number of hyperspheres [88].

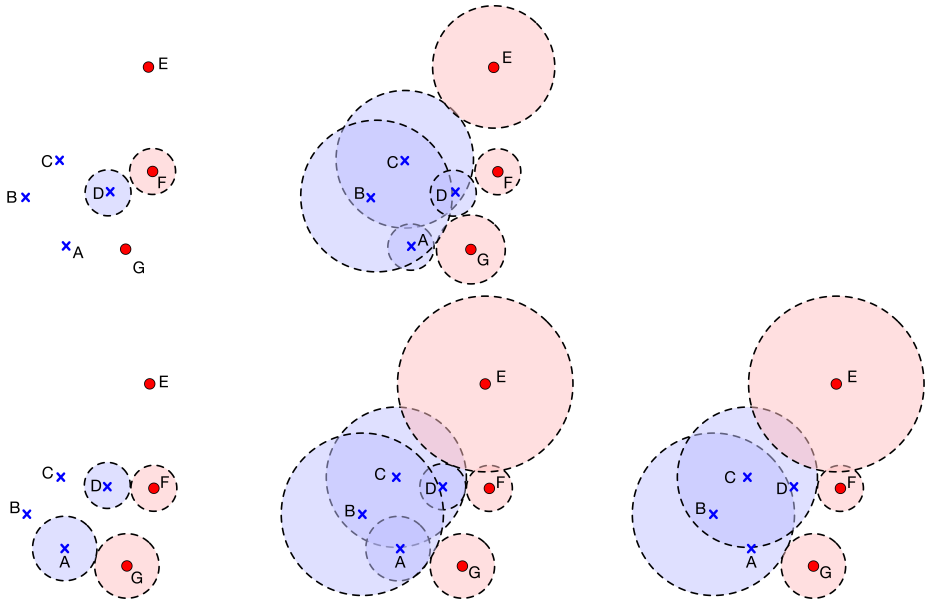

Fig. 10: Alternative T1 implementation [88] for the scenarios depicted in Figure 9. The modification starts by finding which data points are each other's nearest neighbours of opposite classes (i.e., nearest enemies): D and F in the scenario of the top row, and both D and F and A and G in the bottom row. The radii of their hyperspheres are automatically defined as half of the distance between them. Then, for each remaining data point, its radius is defined as the distance to its nearest enemy minus the radius of the nearest enemy itself. Considering the scenario in the top row, the radius of hypersphere C corresponds to its distance to F (its nearest enemy), minus the radius of F itself. Accordingly, the radius of E is determined by considering its distance to C, and so forth.

6.2.3 Local Set Average Cardinality (LSCAvg)

The Local Set (LS) of a given data example x_i is the set of examples whose distance to x_i is smaller than the distance of x_i to its nearest neighbour of the opposite class, NN_{io} [82]. An example of a LS is depicted in Figure 11. Considering U as the set of all examples in the data space, the LS of a given example x_i can be defined as:

$$LS(x_i) = \{x_j \in U : d(x_i, x_j) < d(x_i, NN_{io})\} \tag{12}$$

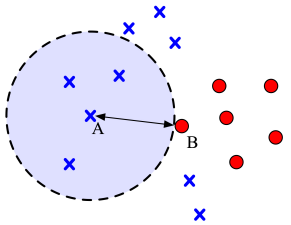

Fig. 11: The concept of Local Set, LS. Considering x_i as point A, its nearest neighbour of the opposite class NN_{io} (nearest enemy) is point B. Thus, the LS of point A is the set of examples whose distance to A is smaller than $d(A, B)$, included in the dotted circle. The local set cardinality of A is therefore 4, i.e., $|LS(A)| = 4$. Adapted from [82].

To determine the Local Set Average Cardinality (LSCAvg) of a dataset, the number of points included in each example's LS is aggregated according to Equation 13. Examples with a small number of points in their LS are either examples located near narrow decision borders, or examples located in regions populated by the opposite class (overlapping regions). A smaller number of points in each example's LS leads to lower values of LSCAvg, which represent more overlapped and complex domains.

$$LSCAvg = \frac{1}{n^2} \sum_{i=1}^{n} |LS(x_i)| \tag{13}$$

, where n represents the total number of examples in data.

6.2.4 Number of Clusters (Clst)

The Number of Clusters ($Clst$), similarly to T1, determines the number of clusters of the same class that cover the data domain [82]. The proposed algorithm in [82] starts by considering the data examples with higher LS cardinality as cluster cores. Then, for each remaining example, the algorithm checks if they belong to the LS of a cluster core. If so, the example is included in the existing cluster; otherwise, a new cluster core is created, and the process is repeated, always prioritising cores with the highest LS cardinality. An example of the clustering procedure is depicted in Figure 12. After all examples are assigned to clusters, the total number of existing clusters is determined and $Clst$ defined by Equation 14, where n is the total number of examples in data.

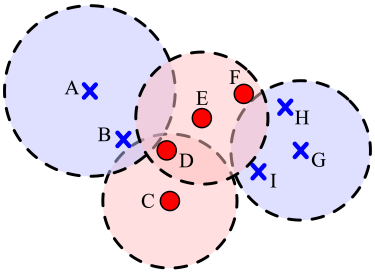

Fig. 12: Local Set-based clustering. The first identified cores are E and G, in any order, since they have the largest LS ($|LS(E)| = |LS(G)| = 3$). Then, points A and C are chosen as cores since they both have a LS of 2. The remaining examples do not become cores, since they are already comprised in the local sets of other cores. Finally, although D is both contained in the LS of E and C, it belongs to the cluster with core E, since E has a higher LS cardinality. Adapted from [82].

$$Clst = \frac{\#Clusters}{n} \tag{14}$$

A note worth considering is that, in the original formulation, LSCAvg and $Clst$ mainly focus on characterising class borders (determining how narrow and/or irregular they are). For this reason, overlapping and noisy examples are considered atypical and removed from the dataset (using the ENN algorithm [144]) prior to the computation of the LSC cardinality of each example. Nevertheless, both types of examples (located near the class borders, or in overlapping regions) contribute to class overlap, and both LSCAvg and $Clst$ can be used to characterise it. Figures 13 and 14 provide an comparison between a solution that does not remove overlapping points and one that does (as originally formulated).

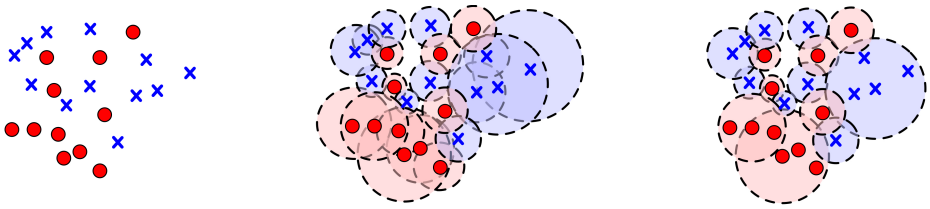

Fig. 13: A representation of the $Clst$ solution for a given dataset, considering all points. The LS of each data example is determined and starting with the examples with largest LS, the clusters are built by iteratively finding candidate cluster cores. In this solution, all existing points are kept and the final number of clusters reflects the amount of class overlap in the domain: 15 clusters for 23 data points.

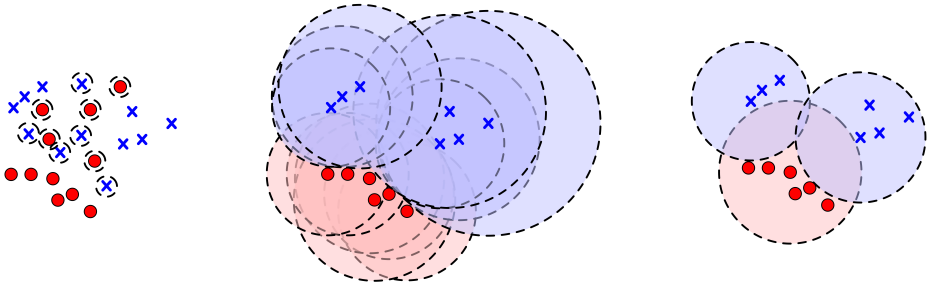

Fig. 14: A representation of the *Clst* solution for the dataset in Figure 13, removing overlapped and noisy points. In this scenario, prior to the LS computation, the noisy and overlapped points are removed according to the ENN rule, returning a solution of 3 clusters for 13 data points. It seems, however, that removing data points alters the true complexity of the original data domain.

6.2.5 Overlap Number of Balls (ONB)

The Overlap Number of Balls (ONB) is based on the same rationale as T1 [105]. The idea is to determine how many balls containing only examples of the same class are needed to cover the entire data space. ONB uses the Pure Class Cover Catch Digraph [94] to determine the maximum radii for all examples in data (the radius of a ball is increased until it touches an example from the opposite class). Then, for each example, the ball that includes the largest number of same-class examples is chosen, until all examples are covered. After the final number of balls is defined, two measures can be determined: ONB_{tot} and ONB_{avg}. ONB_{tot} represents the ratio between the number of balls necessary to cover the domain and the number of examples in data, n (Equation 15). ONB_{avg} determines the average ONB, considering the number of balls necessary to cover each class C (Equation 16).

$$ONB_{tot} = \frac{\text{Number of Balls}}{n} \tag{15}$$

$$ONB_{avg} = \frac{1}{C} \cdot \sum_{c=1}^{C} \frac{\text{balls}_c}{n_c} \tag{16}$$

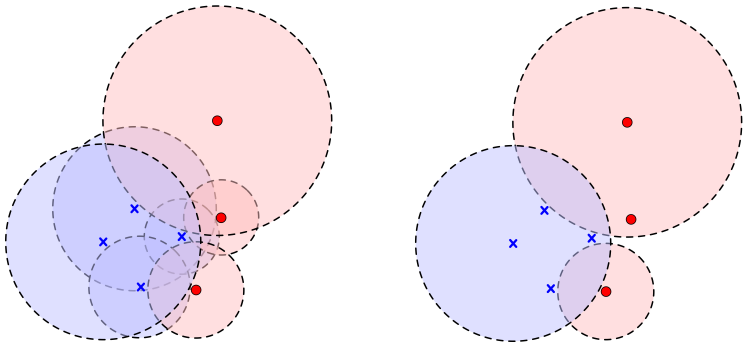

Fig. 15: A representation of the ONB solution for the dataset in Figures 9 and 10 (top-row). First, a ball is centred at each data point and grown until it touches a point from the opposite class. Then, the balls containing a larger number of points are iteratively chosen. Adapted from [105].

6.2.6 Decision Boundary Complexity (DBC)

The Decision Boundary Complexity (DBC) is an extension of T1 which determines the interleaving of hyperspheres of different classes [137]. After the hyperspheres from T1 are found, a Minimum Spanning Tree (MST) is constructed using the centres of the hyperspheres. Then, the number of connected centres of different classes (N_{inter}) is determined and DBC is computed as follows:

$$\text{DBC} = \frac{N_{inter}}{\#Hyperspheres} \tag{17}$$

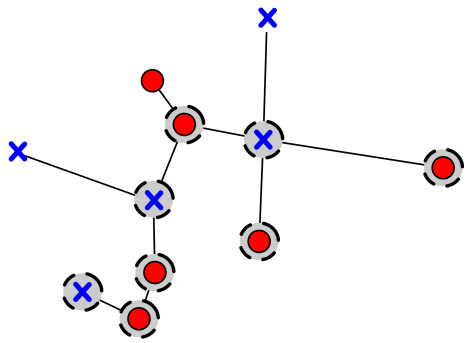

Fig. 16: A representation of the DBC measure. In the MST, there are 8 centres connected to centres of a different class ($N_{inter} = 8$).

6.2.7 Ratio of Intra/Extra Class Nearest Neighbour Distance (N2)

N2 compares the within-class and between-class spread, i.e., it represents a trade-off between intra-class distances and inter-class distances [66]. The distance be-

tween each data example and its nearest neighbour of the same class, $d(x_i, NN_{is})$, as well as between its nearest neighbour from the opposite class, $d(x_i, NN_{io})$, are computed. Then, the sum of all intra and inter-class distances are aggregated to produce a intra/inter class ratio (r) and N2 can be determined according to Equation 18, where n represents the total number of examples in data. Higher values of N2 indicate more overlapped domains [111].

$$\text{N2} = \frac{r}{1+r}, \text{ where } r = \frac{\sum_{i=1}^{n} d(x_i, NN_{is})}{\sum_{i=1}^{n} d(x_i, NN_{io})} \tag{18}$$

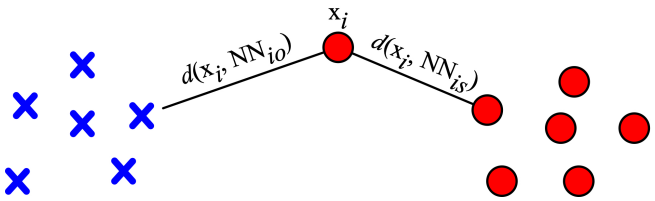

Fig. 17: A representation of intra-distances and inter-distances for N2 computation. Less overlapped domains generally present more compact concepts (lower intra-distances average) that are well-separated (higher inter-distances average), thus returning lower values of N2.

6.2.8 Number of samples per group (NSG)

This measure provides an indication of the average size of groups that exist in data by determining the average number of examples in each hypersphere found by T1 (Equation 19) [136]. N_i represents the number of examples inside hypersphere i.

$$NSG = \frac{1}{\#Hyperspheres} \sum_{i=1}^{\#\text{Hyperspheres}} N_i \tag{19}$$

In such a way, NSG (as all density measures in general) adds local information to structural overlap measures. A large number of hyperspheres comprising a small number of examples is indicative of a more intertwined data domain.

6.2.9 Inter-Class Scale Variation (ICSV)

The inter-class scale variation measures the standard deviation of hyperspheres' densities [136]. First, the density ρ of each hypersphere found according to T1 is determined, where N_{sphere} and V_{sphere} represent the number of examples in a hypersphere and its volume, respectively. Then, the standard deviation of the sphere densities (ICSV) is found, as follows from Equation 20. n_H represents the number of hyperspheres (#Hyperspheres) and μ_ρ represents the average density of hyperspheres. Higher ICSV values are associated with changes in the local data densities of the domain, thus indicating more complex scenarios.

$$ICSV = \sqrt{\frac{1}{n_H}\sum_{i=1}^{n_H}(\rho_i - \mu_\rho)^2}, \text{ where } \rho = \frac{N_{sphere}}{V_{sphere}} \text{ and } \mu_\rho = \frac{1}{n_H}\sum_{i=1}^{n_H}\rho_i \quad (20)$$

6.3 Instance-Level Overlap

These measures are able to analyse the domains at a local level, where class overlap is commonly associated to the error of the k-nearest neighbour classifier. While some measures provide an overall value for the entire domain (R-value, R_{aug}, *degOver*, N3, SI, N4), others are particularly related to the identification of local data characteristics, i.e., data typology or instance hardness (kDN, D3, Borderline Examples, IPoints). They provide local information on the complexity of the domain by identifying problematic examples in data, frequently those near the class boundaries (associated with class overlap). Although some of these measures evaluate data examples individually according to their characteristics, they can then adapted in order to produce an estimate for the entire domain.

6.3.1 R-value and Augmented R-value

The R-value defines the degree of overlap between two classes C_i and C_j by determining the number of points of each class that fall onto overlap regions between classes [103]. For each m^{th} instance of class C_i (represented as p_{im}), the examples in its k-neighbourhood that belong to C_j, represented by $\text{kNN}(p_{im}, C_j)$, are found (Figure 18). Then, p_{im} is assigned as belonging to an overlapping or non-overlapping region, as follows from Equation 21. $|C_i|$ represents the number of examples of class C_i, whereas θ is a threshold used to define whether p_{im} is inside an overlap region or not. λ is a binary function that represents such decision, i.e., $\lambda(a) = 1$ if $a > 0$; otherwise $\lambda(a) = 0$. In other words, if we consider $\theta = 2$, it means that 2 is the maximum number of points from the opposite class that we tolerate in the k-vicinity of p_{im}. If there are more than 2 points, then p_{im} is considered an overlapping point. The same is performed for class C_j and the final results are aggregated as follows from Equation 22.

$$r(C_i, C_j) = \sum_{m=1}^{|C_i|} \lambda\Big(|\text{kNN}(p_{im}, C_j)| - \theta\Big) \quad (21)$$

$$R(C_i, C_j) = \frac{r(C_i, C_j) + r(C_j, C_i)}{|C_i| + |C_j|} \quad (22)$$

R-values range from 0 (no overlap) to 1 (complete overlap), taking into account all examples in the data domain, whether they are from the majority or minority classes.

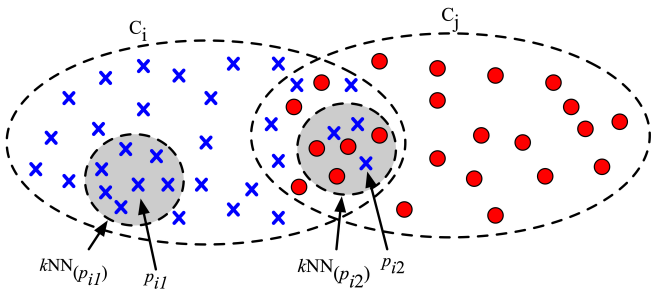

Fig. 18: Basic concepts for R-value computation. Note how $|\text{kNN}(p_{i1}, C_j)| = 0$ and $|\text{kNN}(p_{i2}, C_j)| = 4$, for $k = 6$. Adapted from [103].

The Augmented R-value (R_{aug}) is an extension of the R-value that takes into account the imbalance ratio of the data domain [14] (Equation 23), where $R(C_{min})$ and $R(C_{maj})$ may be calculated as an arbitrary $R(C_i)$ according to Equation 24.

$$R_{aug}(C_{min}, C_{maj}) = \frac{1}{\text{IR} + 1}\Big(R(C_{maj}) + \text{IR} \cdot R(C_{min})\Big) \tag{23}$$

$$R(C_i) = \frac{1}{|C_i|}\sum_{m=1}^{|C_i|}\lambda\Big(|\text{kNN}(p_{im}, C_j)| - \theta\Big) \tag{24}$$

This extension is based on the rationale that, for binary classification, the contribution of the majority class overlap to the overall overlap should not be directly proportional to the number of majority examples, given that most of them are frequently non-overlapping examples [14]. For IR = 1, R_{aug} is equivalent to the R-value (Equation 22), whereas as the IR increases, R_{aug} becomes closer to the R-value of the minority class (Equation 24, assuming C_i as C_{min}).

6.3.2 degOver

Similarly to what was described in the previous section, *degOver* determines the degree of overlap by finding overlapping and non-overlapping examples in a k-neighbourhood ($k = 5$) [98]. For a given example, if all its 5-nearest neighbours are from the same class, then the example belongs to a non-overlapping region (Figure 19). Otherwise, it is considered an overlapping example. Then, the number of overlapping examples (of both classes), i.e., $n_{min_{over}}$ and $n_{maj_{over}}$ is divided by the total of examples in the data space, n (Equation 25). Higher values of *degOver* represent more overlapped domains.

$$\text{degOver} = \frac{(n_{min_{over}} + n_{maj_{over}})}{n} \tag{25}$$

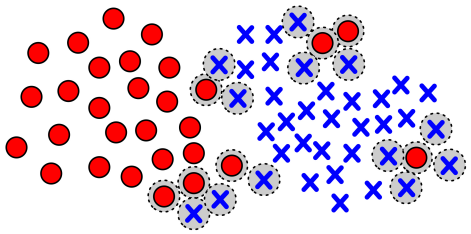

Fig. 19: A representation of *degOver*. Marked points from both classes are those that contribute to class overlap (located in overlapped regions).

6.3.3 Error Rate of the Nearest Neighbour Classifier (N3)

N3 measures the error rate of the Nearest Neighbour classifier (1NN), estimated using a Leave-One-Out (LOO) cross-validation. Higher N3 values are associated with a higher overlap degree between classes [66]. Considering U as the set of all examples in the data space, N3 can be defined according to Equation 26, where y_i represents the class of example x_i, and y_{NN_i} represents the class of its nearest neighbour, NN_i.

$$\text{N3} = \frac{1}{|U|}|\{x_i \in \text{U} : y_i \neq y_{NN_i}\}| \tag{26}$$

6.3.4 Separability Index (SI)

Thornton's Separability Index (SI) determines the proportion of points whose class is the same as of its nearest neighbour [60, 129]. Considering a given example x_i and its nearest neighbour NN_i, SI is defined by Equation 27. In such a way, SI measures class overlap by informing on the separability of the data domain, being the complementary measure of N3, where higher values indicate that there is a large amount of data points whose nearest neighbour is of the opposite class.

$$\text{SI} = \frac{1}{|U|}|\{x_i \in \text{U} : y_i = y_{NN_i}\}| \tag{27}$$

6.3.5 Non-Linearity of the Nearest Neighbour Classifier (N4)

To compute N4, new synthetic examples $\hat{x}_i$ are generated by interpolating pairs of data examples from the same class, chosen randomly. Then, the error rate of the Nearest Neighbour classifier is estimated solely over the set of the new examples obtained by linear interpolation, I. For each new example, its closest neighbour of the original data space NN_{iU} is determined, and their class labels are compared in order to produce N4 (Equation 28). By determining the 1NN error on these new points, N4 establishes the overlap that exists between the convex hulls that delimit the classes [88]. Higher values of N4 represent more deeply overlapped domains.

$$\text{N4} = \frac{1}{|I|}|\{\hat{x}_i \in \text{I} : \hat{y}_i \neq y_{NN_{iU}}\}| \tag{28}$$

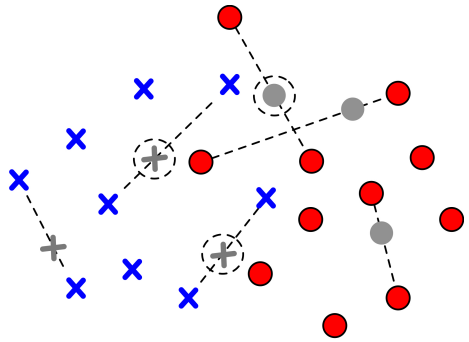

Fig. 20: A representation of N4 computation. New synthetic points (in grey) are generated by linearly interpolating random examples of the same class (connected by dotted lines). Then, the 1NN error is measured over the new points: marked points are those whose 1NN classification produces an error, thus identifying class overlap. Adapted from [88].

6.3.6 K-Disagreeing Neighbours (kDN)

Considering an example x_i, k-Disagreeing Neighbours (kDN) measures the percentage of its k nearest neighbours x_v that do not share its class [120]:

$$\text{kDN}(x_i) = \frac{|\{x_v \in kNN_i : y_v \neq y_i\}|}{k} \tag{29}$$

In such a way, kDN measures the local overlap of a given data example, where values closer to 0 indicate that x_i is inside a safe region (all neighbours share its class label), whereas higher values indicate increasing amounts of data examples from the opposite class in its neighbourhood. A global measure for the entire domain could be achieved by averaging kDN over all examples in data, n:

$$\text{kDN}_{avg} = \frac{1}{n}\sum_{i=1}^{n}\text{kDN}(x_i) \tag{30}$$

6.3.7 Class Density in the Overlap Region (D3)

D3 aims to describe the density of each class in the overlap regions by determining, for each class, the number of examples that lie in regions populated by a different class [122]. For each example x_i, its k-nearest neighbours are found and if the majority belongs to a class different from x_i, then x_i is considered to be in an overlapping region. The number of examples that lie inside overlapping regions is then retrieved for each class C_j. Considering U as the set of all examples in the data space and kNN_i as the set of the k-nearest neighbours of x_i, D3 can be defined according to Equation 31, where higher values for a given class correspond to regions populated by another class. y_i and y_v are the class labels of x_i and x_v, respectively, and Δ_{x_i} establishes the proportion of nearest neighbours of x_i that share its class.

$$\mathrm{D3}_{C_j} = |\{x_i \in U : \Delta(x_i) < 0.5\}| \tag{31}$$

$$\Delta_{x_i} = \frac{|\{x_v \in kNN_i : y_v = y_i\}|}{k} \tag{32}$$

6.3.8 Complexity Metric Based on k-nearest neighbours (CM)

CM also focuses on the local neighbourhood of each example to decide on its difficulty for classification [4]. The k nearest neighbours of each example x_i are found (where k is odd), and if the majority of neighbours is of the same class as x_i, the example is considered easy; otherwise it is considered difficult. CM then measures the proportion of difficult examples in data, as defined in Equation 33, where $\mathrm{kDN}(x_i)$ has been previously described (Equation 29) and n is the total number of examples in data. CM is therefore intrinsically related to kDN and somewhat the aggregation of D3 over the entire domain. Recent extensions of CM include wCM (Weighted Complexity Metric), and dwCM (Dual Weighted Complexity Metric) [116], that use a weighted kNN approach rather than a standard kNN classifier.

$$\mathrm{CM} = \frac{|\{x_i : \mathrm{kDN}(x_i) > 0.5\}|}{n} \tag{33}$$

6.3.9 Borderline Examples

As discussed in Section 3, the presence of borderline examples is closely related to the problem of class overlap since higher percentages of this type of examples complicate the decision boundary between classes. A popular data typology divides data examples into 4 categories [100, 101, 124, 145], according to their local neighbourhood (typically $k = 5$), as follows:

- *Safe* examples have 0 or 1 neighbours of the opposite class;
- *Borderline* examples have 2 or 3 neighbours of the opposite class;
- *Rare* examples have 4 neighbours of the opposite class. Additionally, the only neighbour of the same class should be either an *outlier* example, or a *rare* example as well;
- *Outlier* examples have all 5 neighbours of the opposite class.

A representation of each type of example is presented in Figure 21. Most often, the data typology is used in scenarios comprising class imbalance [100, 101, 124, 145], and therefore is often solely applied to the minority class. However, it can be applied to all existing classes. In such a case, the number of borderline examples from all classes ($n_{borderline}$) is determined according to the rules described above and divided by the total number of examples in data (n), thus defining the degree of overlap as a percentage (Equation 34). This would be reminiscent of R-value, *degOver*, and CM, although it considers solely one type of difficult examples (borderline examples), as they relate the most to the concept of class overlap.

$$\text{Overlap } (\%) = \frac{n_{borderline}}{n} \times 100 \tag{34}$$

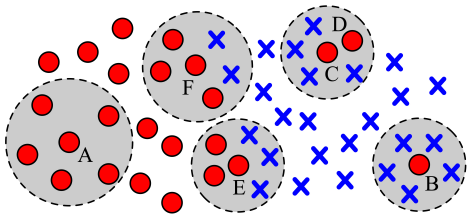

Fig. 21: A representation of different example types: A is a safe example, surrounded by neighbours of its class; B is an outlier example, isolated in an area of the opposite class; C and D are rare examples and finally, E and F are borderline examples, located near the decision border between classes.

6.3.10 Number of Invasive Points (IPoints)

When data examples are clustered according to the their local sets (LS), some resulting clusters may contain only one instance. This may represent a situation where two cluster cores share some examples in their local sets, except than one of the cores has a larger local set cardinality [82]. An example of such situation has previously been discussed in Figure 12, where cores E and C share point D, but D belongs to the cluster with core E, since E has a higher LS cardinality. Then, point C will produce a separate cluster of only one point (itself). If a given cluster has only one point (the core) and its local set contains only the point itself, then it is called an "invasive point". Note that in Figure 12, point C is not an invasive point because, although it will produce a cluster of only itself, its local set contains C and D, i.e., LS(C) = {C,D}. An example of an invasive point is given in Figure 22.

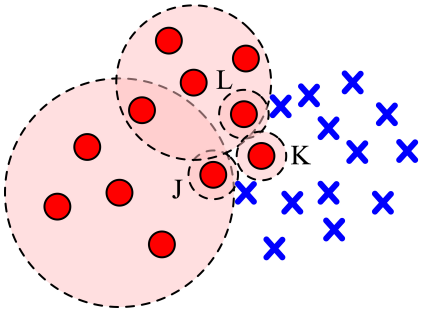

Fig. 22: A representation of an invasive point. Note that K is an invasive point since it produces a cluster of only itself, has no other points in its local set and is not included in the local set of any other point, including its closest neighbours, J and L. In turn, J and L are not invasive points because despite their local sets contain only themselves, they do not produce singular clusters, as they are included in other points' local sets (other clusters). Adapted from [82].

Invasive points are therefore border examples that somewhat infiltrate the opposite class, or examples located in overlapping regions of the data space. The number of these type of points normalised by the total number of points (n) characterises the complexity of the domain, where a large number of invasive points indicates more intertwined domains (Equation 35).

$$\text{IPoints} = \frac{\#Invasive\ Points}{n} \tag{35}$$

6.4 Multiresolution Overlap

This group of measures uses multiresolution approaches to identify regions of different complexity within the domains. Some are more closely related to the previous ideas of using hyperspheres (MRCA) or k-neighbourhoods (C1 and C2) to define regions of the space where class overlap can be analysed. Others are associated with feature space partitioning, where features are divided into a specific number of intervals where the properties of class overlap may be assessed (Purity and Neighbourhood Separability). Nevertheless, the main idea than binds these measures together is that they operate recursively (fine-grain search), i.e., defining hyperspheres, neighbourhoods, or feature partitions at different resolutions, all of which are individually analysed. This allows to combine both local and structural information, characterising the data domains from the perspective of recursive data subspaces. Class overlap is therefore determined at several resolutions, providing a trade-off between global and local data characteristics.

6.4.1 Multiresolution Complexity Analysis (MRCA)

Multiresolution Complexity Analysis (MRCA) aims to identify regions of different complexity in the data domain [5]. Each data example is attributed a *profile space*, which is then used for clustering and complexity analysis. To generate a profile space for a given data example, hyperspheres of different radii are drawn around it. The content of each hypersphere is then analysed through the use of an *imbalance estimation function* which, given a set of examples $\mathbf{D}$, is defined as follows:

$$\psi_D(\mathbf{x}, \sigma) = \mathbf{y}(\mathbf{x}) \cdot \frac{N_\sigma^+(\mathbf{x}) - N_\sigma^-(\mathbf{x})}{N_\sigma^+(\mathbf{x}) + N_\sigma^-(\mathbf{x})} \tag{36}$$

The data example $\mathbf{x}$ and parameter σ are the centre and radius of the hypersphere, respectively, and $N_\sigma^+(\mathbf{x})$ and $N_\sigma^-(\mathbf{x})$ are the number of data examples of the positive and negative class inside the hypersphere. $\mathbf{y}(\mathbf{x})$ gives the class of $\mathbf{x}$, herein assuming two possible values {-1,1}. ψ therefore ranges between [-1,1], where -1 and 1 indicate a strong imbalance inside the hypersphere, with most of the data examples being from the opposite class of $\mathbf{x}$ (-1), or mostly equal to $\mathbf{x}$ (1). $\psi = 0$ characterises situations where both classes are equally represented inside the hypersphere.

A profile pattern of $\mathbf{x}$ can be obtained by considering different radii σ in the generation of the hyperspheres. Considering a set of m hyperspheres, a profile $\mathbf{p}$ is given by:

$$\mathbf{p} = [\psi(\mathbf{x}, \sigma_1), \psi(\mathbf{x}, \sigma_2), \ldots, \psi(\mathbf{x}, \sigma_m)] \tag{37}$$

After all data examples have been assigned their profile patterns, a set of profile patterns Δ is obtained, which can then be clustered to determine regions of different complexity, via k-means clustering [5]. Then, to define the pattern and cluster complexity, a Multiresolution Index (MRI) can be computed for each pattern $\mathbf{p}$:

$$MRI(\mathbf{p}) = \frac{1}{2m} \cdot \sum_{j=1}^{m} w_j \cdot (1 - p_j), \tag{38}$$

where $w_j = 1 - \frac{j-1}{m}$, giving higher weights to components with finer granularity. The complexity of a k^{th} cluster is then determined by averaging the complexity of patterns $\mathbf{p}$ that belong to it.

$$MRI^{(k)} = \frac{1}{|\Delta^{(k)}|} \cdot \sum_{\mathbf{p} \in \Delta^{(k)}} MRI(\mathbf{p}) \tag{39}$$

Lower values of $MRI^{(k)}$ characterise clusters comprising patterns $\mathbf{p}$ with most $\psi_D(\mathbf{x}, \sigma) \approx 1$, which represent patterns $\mathbf{x}$ belonging to less complex regions. In turn, higher values of $MRI^{(k)}$ indicate clusters comprising patterns $\mathbf{p}$ with most $\psi_D(\mathbf{x}, \sigma) \approx -1$, representing patterns $\mathbf{x}$ in more complex regions. Balanced clusters indicate medium complexity regions, with $MRI^{(k)} = \frac{1}{2}$.

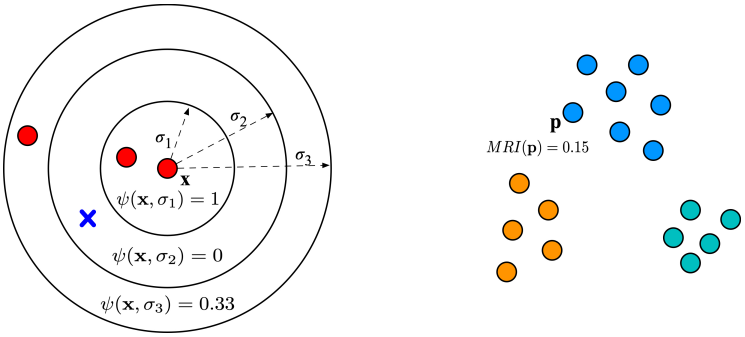

Fig. 23: A representation of MRCA. The profile of data example $\mathbf{x}$ is defined using 3 hyperspheres of radius σ_1, σ_2 and σ_3, for which $\psi(\mathbf{x}, \sigma)$ is computed, respectively. Thus, a profile pattern $\mathbf{p}$ is constructed as $\mathbf{p} = [1, 0, 0.33]$, with a MRI($\mathbf{p}$) of 0.15. After all data examples have been profiled, a new data space of profile patterns Δ is constructed and clustered, where each pattern $\mathbf{p}$ is included in clusters of different complexity. Data example $\mathbf{x}$ was mapped to a pattern $\mathbf{p}$ that belongs to the blue cluster. In such a way, it is possible to find patterns $\mathbf{p}$ of different difficulty by analysing the cluster solution, which in turn correspond to difficult data examples $\mathbf{x}$ in the original data space. Note that patterns $\mathbf{p}$ included in the same cluster do not necessarily correspond to nearby examples in the original data space since clusters are built based on the difficulty of data examples, not their distance to each other.

6.4.2 Case Base Complexity Profile (C_1)

Similarly to MRCA, C_1 measures the local complexity of a data domain by focusing on the spatial distribution of data examples [95]. The complexity of each data example is determined based on the class distribution within its k-neighbourhood, for increasing values of k. For each k value and data example x_j, the proportion of examples that share the same class as x_j is determined (p_{kj}) and a nearest neighbour profile can be determined by plotting p_{kj} as a function of k (Figure 24).

For a given chosen K, the complexity of x_j is given by Equation 40, where neighbours closer to x_j have a higher influence on the complexity since they are used to compute several values of p_{kj} [31].

$$Complexity(x_j) = 1 - \frac{1}{K}\sum_{k=1}^{K} p_{kj} \tag{40}$$

To provide an overall complexity value for the entire data domain, the complexity of all points may be averaged according to Equation 41, where n is the total number of data examples.

$$C_1 = \frac{1}{n}\sum_{j=1}^{n} Complexity(x_j) \tag{41}$$

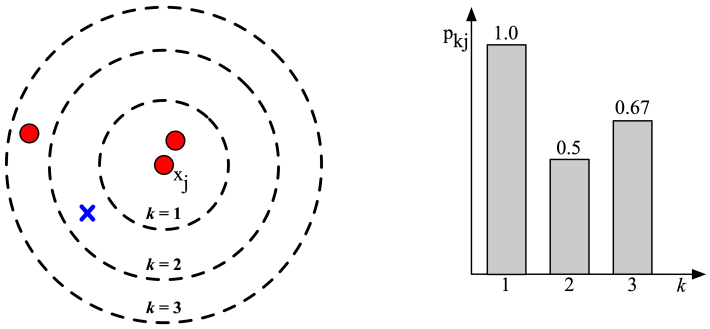

Fig. 24: A representation of C_1. A complexity profile can be determined for x_j by analysing the characteristics of its neighbourhood for different values of k. With $k = 3$, the complexity of x_j is $1 - \frac{1}{3}(1 + 0.5 + 0.67) \approx 0.28$. Adapted from [95].

6.4.3 Similarity-Weighted Case Base Complexity Profile (C_2)

C_2 is a modification of C_1 that associates the weight of each neighbour to their distance to x_j, so that closer neighbours have a higher impact in complexity computation [31]. In C_2, p_{kj} is given as the average similarity between x_j and the k-neighbours that share its class. The overall complexity C_2 is given by the same Equations 40 and 41, yet considering the modifications to p_{kj}.

6.4.4 Purity and Neighbourhood Separability

Another type of multiresolution analysis is feature space partitioning. Feature Space Partitioning measures work by recursively partitioning the data space into hypercuboids (cells) at several resolutions, where each resolution is defined by the number of partitions per feature [117, 118]. As the resolution increases, the data space is composed by a larger number of cells and each cell includes a smaller number of data examples. Based on this partitioning scheme, two complexity measures called *Purity* and *Neighbourhood Separability* may be defined. The former relates to how pure are the defined cells, considering the number of representatives of each class comprised inside each cell. The latter finds, for each example in a cell, the proportion of nearest neighbours that share its class.

For both measures, the data space is divided at different resolutions from $B = 0$ (no partitioning) to $B = 31$ (up to 32 cells per axis), where data examples are assigned to their closest cell (Figure 25). Then, the following strategy is applied:

- At each resolution B, the complexity (*purity* or *neighbourhood separability*) is measured individually for each cell;
- The estimates of each cell are linearly weighted to produce an estimate for that resolution, where the weight given to the estimate of each cell is proportional to the number of examples it contains ($\frac{n_l}{n}$), where n_l is the number of examples in the cell and n represents the total number of examples in data;
- The complexity across all cells at a given resolution is also exponentially weighted by a factor of $w = \frac{1}{2^B}$, where larger weights are given to lower resolutions;
- Finally, a curve of complexity versus resolution is plotted and the area under the curve (AUC) defines the overall complexity of the data, bounded within the [0,1] interval.

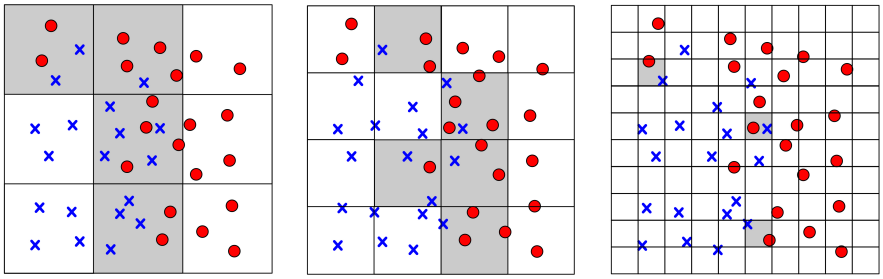

Fig. 25: A representation of the feature partitioning scheme for $B = 2$, $B = 3$ and $B = 9$, from left to right, respectively. Higher resolutions provide more local information regarding the domain. At each resolution B, the domain complexity (*purity* or *neighbourhood separability*) is determined, where each cell is individually analysed. The final complexity measures are determined by averaging the individual results of all cells. Cells marked in grey are those shared by examples of different classes, identifying overlapping regions.

In what follows, we explain how *purity* and *neighbourhood separability* are computed. Detailed algorithms of both measures, as well as the feature partitioning scheme, are available in [118].

Purity

Purity measure determines how pure the defined cells are, focusing on class representation inside each cell. If all data examples are from the same class, the cell is completely pure; otherwise, the purity of each cell depends on the number of representatives of each class comprised inside it. In the worst case, if a cell contains the same number of examples for each class, its purity is zero.

Considering a total of K_l classes in cell H_l, and considering that the number of examples of class C_i in cell H_l is given by λ_{il}, the purity of a cell is defined as:

$$S_{H_l} = \sqrt{\left(\frac{K_l}{K_l - 1}\right) \sum_{i=1}^{K_l} \left(p_{il} - \frac{1}{K_l}\right)^2} \tag{42}$$

where p_{il} is the probability of class C_i in H_l, given by:

$$p_{il} = \frac{\lambda_{il}}{\sum_{i=1}^{K_l} \lambda_{il}} \tag{43}$$

The estimates S_{H_l} of each cell are then linearly weighted and summed to produce an average purity S_H, given by:

$$S_H = \sum_{l=1}^{H} S_{H_l} \cdot \frac{n_l}{n} \tag{44}$$

where H is the total number of cells. A previously detailed, S_H is further weighted by $\frac{1}{2^B}$ before plotting the purity values versus resolution. The overall purity measure, i.e., the AUC of purity values across all cells (S_H) versus the respective resolution (B), is bounded within the range [0,1] where higher values represent less overlapped domains. For less overlapped domains, the purity is expected to increase as the number of cells increases with higher resolutions. However, if the domain is extremely overlapped, the purity will be low despite the increase of the number of cells, therefore returning a lower average purity value. Additionally, for less overlapped domains, the measure will increase rapidly as the resolution increases, contrary to data with significant class overlap.

Neighbourhood Separability

This measure is more sensitive to the shape of decision boundaries and determines, for each data example in a cell, its proportion of k-nearest neighbours from the same class (for varying values of k). For each data example x_j in cell H_l, its k-nearest neighbours are found based on the Euclidean distance and the proportion of neighbours from the same class as x_j is determined as p_{kj}. This procedure is repeated for several values of k, from 1 to a maximum value of λ_{il}, in steps of 1 (recall that λ_{il} is the number of examples of class C_i inside cell H_l). Thus, for each data example x_j inside cell H_l, it is possible to plot a curve of p_{kj} versus k

and determine the area under the curve as ϕ_j. Then, the average neighbourhood separability of cell H_l can be determined as:

$$p_l = \frac{1}{n_l} \sum_{j=1}^{n_l} \phi_j \tag{45}$$

The neighbourhood separability across all cells is computed by a weighted sum of the p_l values of all cells (Equation 46) and then weighted by $\frac{1}{2^B}$ to account for the data space resolution.

$$S_{NN} = \sum_{l=1}^{H} p_l \cdot \frac{n_l}{n} \tag{46}$$

Similarly to *Purity*, a final curve of S_{NN} values versus resolution is plotted and the area under the curve is the overall neighbourhood separability measure for a given domain, where higher values represent less overlapped domains.

6.5 Summarizing Comments

Throughout this section we discuss the idea that class overlap is a heterogeneous problem with different representations. To standardise existing vortices of class overlap, we propose a novel taxonomy that associates common concepts found in related research to four groups of class overlap complexity measures (Figure 4): Feature Overlap, Structural Overlap, Instance-Level Overlap, and Multiresolution Overlap. We show how each group measures a particular facet of class overlap and describe their representative measures in detail, which is a step towards providing a more complete characterisation of class overlap in real-world domains. However, there are two topics left for discussion. One is if (and how) these measures of class overlap are attentive to class imbalance as well. The other regards the development of new measures that simultaneously account for several representations of class overlap. Let us start by discussing the existing body of knowledge regarding the sensitivity of class overlap measures to class imbalance.

As highlighted in Figure 4, there are some measures for which adaptations to imbalanced domains are discussed in the literature. Some were originally developed in the scope of imbalanced data (R_{aug}, ONB, CM, dCM, dwCM, Borderline Examples), while others correspond to the recently-suggested, class-wise adaptations of well-known complexity measures (F2, F3, F4, N1, N2, N3, N4, T1) [9]. The underlying motivation for these adaptations is that, since certain measures consider classes altogether, the majority class tends to dominate their computation and hence they perform poorly in imbalanced domains [4, 8, 143]. Current adaptations are therefore based on evaluating the individual class complexities, i.e., decomposing measures into their minority and majority counterparts. As an example, consider the original N3 measure which determines the error of a 1NN classifier. The adapted version of N3 consists of taking the 1NN error per class. The adapted measures have shown promising results in estimating the difficulty of classification tasks more accurately than the original measures for binary-classification domains [8, 9], although this is still a line of ongoing research. Except for the measures discussed herein (and marked in Figure 4), there are no considerations

regarding the remaining in what concerns imbalanced domains. Naturally, in the same light of the results previously reported, we can expect a biased behaviour for certain measures (e.g., those that provide average values over the total number of examples in data). Nevertheless, others require further investigation.

The devise of adaptations and combinations of existing representations (*ergo*, measures) of class overlap remains an open challenge for future research. Although the presented taxonomy is insightful to associate existing measures to different class overlap representations, each group of measures still gives emphasis to a particular facet. To provide a complete characterisation of the problem of class overlap for a given domain and a full understanding of to what extent it is harming the classification task, it is required that these measures are either used collectively or combined to capture several representations simultaneously. The idea that, in imbalanced domains, class overlap may be more thoroughly characterised by measures that consider multiple sources of complexity is recently touched upon in [105]. With the development of ONB, authors explore the suitability of combining structural, local, and class imbalance information to provide good estimates for class overlap.

Although both topics are currently under research, they show that there is somewhat a consensus in what concerns the limitations of individual measures of class overlap, and the need to characterise the class overlap problem in all its dimensions, while also accounting for class imbalance. This is one of the biggest open challenges in the imbalanced data field and the reason why a unified view on the problem is necessary to put forward.

In the next section, we will review the state-of-the-art class overlap-based approaches applied to real-world imbalanced domains. We will show that, although under the same rationale of minimising class overlap, the methods often approach the problem from different perspectives, i.e., focusing on different representations of class overlap. Also, despite the fact that several class overlap measures have been discussed in the literature, related research often fails to characterise the problem in the domains, which complicates the evaluation of the efficiency of the approaches, besides preventing the generation of informed recommendations for researchers.

7 Class Overlap-Based Approaches

The topic of learning from imbalanced data has been extensively studied in the past years, with several outstanding survey papers being recently published [34, 46, 62, 72, 76]. As such, the characterisation of the problem of class imbalance and respective taxonomies of approaches and applications is quite well-established. However, few works have attempted to provide a global view on the problem of class overlap in imbalanced domains that summarises, categorises and compares the state-of-the-art strategies used to handle both problems simultaneously. Xiong et al. [147] suggest that data in overlapping regions can be handled by *discarding*, *merging*, and *separating* schemes. In brief, the *discarding* scheme only learns from non-overlapping regions, disregarding the remaining. The *merging* scheme considers the overlapped data as a new class, whereas the *separating* scheme treats overlapping and non-overlapping regions separately, i.e., two separate models are built for each scenario. Most recently, Pattaramon et al. [135] divide class overlap

methods depending on whether methods address all overlapping examples or just those closer to the decision boundaries (borderline examples).

Nevertheless, the relationship between existing class overlap approaches and class overlap representations remains somewhat hidden. This naturally hinders the devise of recommendations for researchers, i.e., it is not possible to determine which approaches would be best for a given domain based on its characterisation. Ultimately, this would be a game-changing contribution to research: guide the choice of appropriate methods or the development of specialised approaches based on the characteristics of the domains, going towards a meta-learning logic. Throughout this section, we will show that, unfortunately, this remains an open issue due to certain limitations found in current research, which will be summarised at the end of this section. However, we thoroughly analysed the existing class overlap-based approaches in order to associate their internal behaviour to the characteristics of data they are sensitive to. With that, we propose a novel taxonomy of class overlap-based approaches aligned with the taxonomy of class overlap complexity measures presented in the previous section.

Figure 26 depicts the most common approaches to handle imbalanced and overlapped domains, together with the class overlap representations, information, and concepts they are associated to. In imbalance data learning, resampling approaches – undersampling and oversampling – are by far the most popular [111]: it comes therefore at no surprise that they remain two of the most explored approaches when handling class imbalance and overlap simultaneously. In addition, cleaning approaches are also frequently applied, either alone or in combination with undersampling and oversampling. Finally, recent research has also explored the use of ensembles, region splitting, evolutionary, and hybrid approaches. In what follows, we describe the proposed taxonomy in higher detail, illustrating each category with both well-established and emergent approaches studied in the context of imbalanced and overlapped domains. To help the reader navigate this section, Table 2 provides an overview of the discussed class overlap-based approaches. Each approach is characterised in what concerns its category (according to the established taxonomy) and the type of information it relies on. The measures used to characterise the data domains in what concerns class imbalance and overlap, as well as the benchmark of compared approaches used in the respective research work are also presented.

7.1 Undersampling Approaches

Undersampling approaches focus on removing redundant majority examples from data and often involve the application of cluster-based methods, thus taking advantage of structural overlap information to identify and characterise overlapping regions in the domain. Based on the internal behaviour of methods proposed in related research, we further divided cluster-based methods into three main types: density-based, neighbourhood-based and fuzzy-based approaches.

Density-based approaches make use of information regarding the density of manifolds to define clusters in data and often rely on the well-known DBSCAN algorithm [40]. A recent example is **ClusBUS** [32], which discards majority examples lying on overlapping regions by using DBSCAN to find clusters that contain both minority and majority examples, and removing enough majority examples to

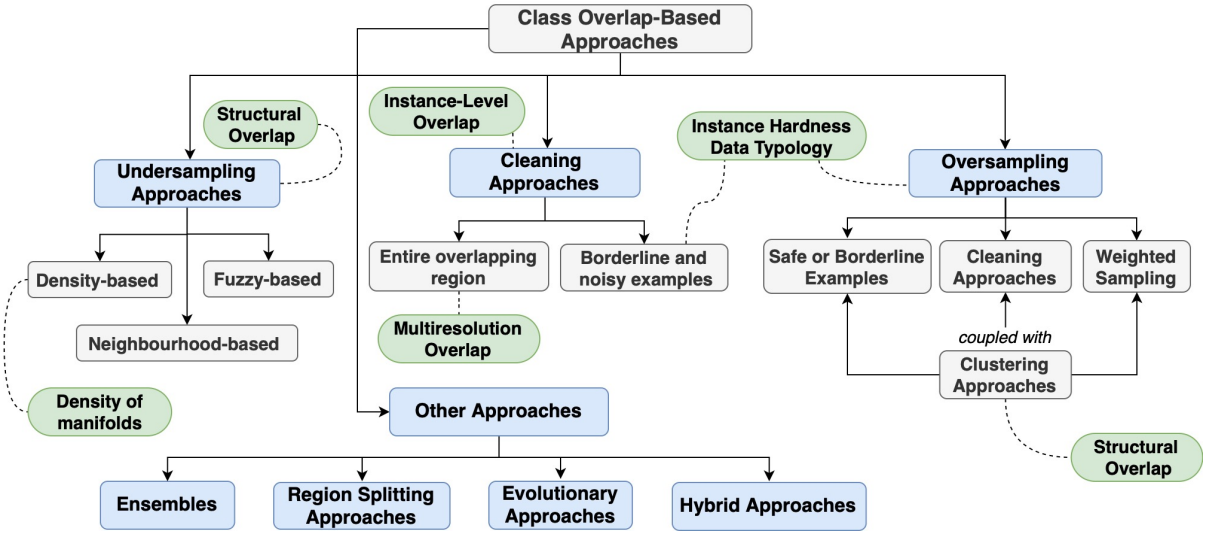

Fig. 26: A taxonomy of methods for handling imbalanced and overlapped datasets. The scheme shows the different class overlap-based approaches that are analysed in this section, associating each group to common class overlap concepts and representations found in related research.

define a vacuum region surrounding minority examples. As previously discussed in Section 6, structural overlap measures may observe a combination of both geometrical and graph-based properties (e.g., hypersphere coverage and MST), and include measures of data sparsity and density of manifolds. Similarly, density-based undersampling algorithms often incorporate both density-based and graph-based procedures. **DBMUTE** uses DBSCAN to define a blemished graph and eliminate majority examples from the overlap region [16]. **DBMIST-US** handles overlapping and noisy majority examples through a combination of DBSCAN clustering with a minimum spanning tree [61].

When the clustering algorithm is k-means, the undersampling approaches rely mostly on neighbourhood-based information (distances between examples). In the context of imbalanced and overlapped domains, k-means is used to define the major core concepts in data, whereas complicated or redundant examples are further removed from the training set. **ClusterOSS** [7] is an extension of OSS (One-Sided-Selection [77]) that uses k-means to choose the candidate majority examples to start the OSS algorithm. Afterwards, borderline and noisy majority examples are removed using Tomek links [130]. In turn, **CUST** first removes borderline majority examples using Tomek links and the remaining redundant and noisy majority examples are eliminated after k-means analysis [123].

Finally, some approaches consider soft-clustering algorithms to look for (and eliminate) overlapping majority examples. This is the case of **OBU**, which uses Fuzzy C-means to establish class-membership degrees to majority data examples [134]. Indecisive examples (those with unclear membership) are considered to be overlapped and are therefore removed. **AdaOBU** further incorporates an

adaptive elimination threshold in OBU allowing its generalisation to datasets with varying overlap degrees [132].

7.2 Cleaning Approaches

Cleaning approaches focus on cleaning the training set by eliminating redundant and/or harmful examples for classification. They may remove examples only from the majority or minority classes, or both (in a two-classification problem). In imbalanced and overlapped domains, however, cleaning approaches are often used as undersampling approaches, since the eliminated examples are often exclusively from the majority class.

All cleaning approaches consider local information, i.e., they commonly rely on instance-level overlap. Some focus on cleaning complicated examples near the decision boundaries, thus analysing local data characteristics (data typology or instance hardness). Accordingly, they determine the safeness level of individual examples to define which should be removed (e.g., evaluating 1NN rules, kDN rules or searching for borderline examples). Others offer a more deep cleaning throughout the entire domain, handling examples that may be located far from the class borders.

Let us start with more seminal cleaning approaches, which were traditionally conceived to eliminate harmful examples irrespective of their class, and focused mostly on borderline examples. **Tomek Links (TL)** [130] define a pair of examples from different classes that are each other's closest neighbours and can be used as a cleaning approach (removing both points) or undersampling approach (removing just the majority point). The **Condensed Nearest Neighbour Rule (CNN)** [64] eliminates redundant examples by keeping only a consistent subset of examples, i.e., those from which a 1-nearest neighbour rule would be able to correctly classify the remaining. Similarly, CNN can be used as an undersampling approach (US-CNN) by keeping all minority examples and producing a subset of majority examples. The **One-Sided-Selection (OSS)** technique [77]) can alleviate the problem of class overlap in imbalanced domains by combining the US-CNN and the concept of TL to remove redundant, borderline, and noisy majority class examples in data. The **Edited Nearest Neighbour (ENN)** rule [144] removes data examples that are misclassified by their k-nearest neighbours (typically $k = 3$). It can be used as an undersampling method by eliminating only majority class examples. Similarly, the **Majority Undersampling Technique (MUTE)** [18] eliminates majority examples whose k-neighbourhood is entirely from the minority class and can therefore be considered a cleaning approach as well. Finally, another well-known cleaning approach is the **Neighbourhood Cleaning Rule (NCL)** [80], which is similar to OSS, although it emphasises more the data cleaning procedure by using ENN. These are some well-established cleaning approaches that can be used as (or incorporated in) undersampling approaches, or even coupled with oversampling approaches (e.g., SMOTE-TL and SMOTE-ENN [11]). Cleaning procedures have proven to enhance classification results by removing overlapped examples that existed in the original training dataset or created during the synthetisation of new examples [111].

Overall, the above approaches aim to clean complicated examples near the class boundaries, therefore focusing mostly on borderline regions. However, as discussed

throughout the paper, despite the fact that borderline examples are a frequent representation of class overlap, there are other types of examples scattered throughout the domain that also contribute to class overlap. Most recently, Pattaramon et al. [133] proposed a set of cleaning approaches (used for undersampling) that focus on providing a deeper level of elimination of harmful examples. They are all based on neighbourhood analysis (instance-level overlap) and therefore identified with the NB- (i.e., "neighbourhood based") prefix. The **Basic Neighbourhood Search (NB-Basic)** removes any majority example that has a minority neighbour. The **Modified Tomek Link Search (NB-Tomek)** removes any majority example with a minority neighbour, only if it appears within the k-neighbourhood of that minority example. In the **Common Nearest Neighbours Search (NB-Comm)**, the common majority nearest neighbours of any two minority examples are identified as overlapped examples and removed. Finally, the **Recursive Search (NB-Rec)** combines local information with multiresolution (fine grain search) information. It starts with the majority examples to be eliminated by NB-Comm and uses them as secondary queries for NB-Rec. The majority examples that are the common nearest neighbours of any pair of these secondary queries are then eliminated as well. By introducing this extension, a finer grain-search criteria is provided and as a result, a higher number of overlapped majority examples is detected and removed.

7.3 Oversampling Approaches

Oversampling approaches generally focus on generating new minority examples to mitigate the problem of class imbalance. In overlapped domains, the main concern of oversampling approaches is to increase the representation of minority examples in specific regions of the data space. For that reason, they often rely on instance-level overlap (local information) to look for candidate examples to guide the synthetisation process.

By far, the most well-known oversampling approach is the **Synthetic Minority Oversampling Technique (SMOTE)** [21]. Although it successfully balances the data domain, SMOTE has no particular mechanism to alleviate class overlap and may even generate overlapping examples if the oversampling procedure occurs near the class borders or includes noisy examples located within the majority class (the problem of overgeneralisation [47]). However, over the years, several modifications of SMOTE have been proposed [73], more and more tailored to certain characteristics of the data domain, including class overlap. Some approaches focus either on improving the representation of examples in the borderline regions between classes (**Borderline-SMOTE**), or in safe regions of the data space (**Safe-Level SMOTE**) [17, 63]. Other approaches search the entire domain and give a higher weight to examples that are harder to learn and should therefore be oversampled more often (**ADASYN**) [65]. To do so, they mostly consider instance-hardness and data typology information, namely variations of the kDN measure. Also considering instance-hardness information are the approaches that incorporate cleaning procedures. These often couple SMOTE with some of the cleaning procedures discussed above, namely **SMOTE-ENN**, **SMOTE-TL** [11], and **SMOTE-IPF** [108]. **SPIDER** [126] is another example, which couples oversampling with deletion of noisy examples. In this case, SPIDER also redirects the

oversampling towards either only borderline or both borderline and safe regions, depending on the chosen amplification. Although there are different variations, these approaches are based on the same underlying information that considers the kDN of each minority example to decide on their probability of oversampling and/or their removal from the dataset. Despite these approaches generally improve the performance of classifiers over imbalanced and overlapped domains, they have well-known handicaps [111]. Several SMOTE-like methods, by using the same interpolation as SMOTE, are prone to the same problem of overgeneralisation, and may generate examples in overlapping areas. Also, in some cases, if the probability of examples to be oversampled is the same across the domain, some redundant minority examples might be oversampled unnecessarily. Finally, noisy minority regions can also be oversampled and remain even after the cleaning procedure. These handicaps occur because the above approaches are focused only on analysing local information, disregarding the structure of both minority and majority classes. Thus, recent research is starting to explore approaches that also consider other types of information, namely structural information.

As previously discussed, one popular way to consider structural information of the domain is via clustering approaches. To that regard, **AHC** [28], **CBO** [70], **DBSMOTE** [19], and **MWMOTE** [10] are popular cluster-based approaches that attend simultaneously to structural and instance-level overlap information on the domain. To this regard, MWMOTE has proven to be a strong competitor over traditional oversampling approaches, due to its further ability to aggregate other types of operations (clustering, cleaning, and adaptive weighting of examples) [111].

Similarly, other recent oversampling algorithms are starting to combine different types of information (structural overlap, data typology, and instance hardness) and operations (clustering and cleaning). **ASUWO** synthesises more examples in the sub-clusters with higher misclassification errors [102]. **IA-SUWO** [141] is an extension of ASUWO that considers a different weighting scheme for minority examples (least squares support numerical spectrum values) and the k-information nearest neighbour method in the oversampling stage. **NI-MWMOTE** (a MWMOTE extension) starts by adaptively removing noise [142]. Then, it uses AHC to segment the minority class examples and adaptively determine the number of examples to synthesise in each sub-cluster using misclassification error as a measure of cluster complexity. The oversampling is performed using MWMOTE. An interesting detail of NI-MWMOTE is that it also uses information regarding the density of manifolds (neighbours' density) to distinguish between suspected and real noise. Another example is **PAIO**, which divides the minority examples using a density-based clustering method similar to DBSCAN (NBDOS), and then defines different interpolation strategies for each type of minority examples [152]. In this case, rather than the standard data typology defined by k-neighbourhood analysis, PAIO uses NBDOS to distinguish between *inland* examples, *borderline* examples, and *trapped* examples.

There are also recent approaches where clustering is more aligned with the concept of hypersphere coverage. **CCR** combines cleaning with oversampling by introducing a energy-based ball coverage strategy [74]. Each minority example has an associated sphere and energy budget, and the sphere is expanded until there is no available energy. When the expansion can no longer proceed, the majority examples are pushed out of the spheres (though not eliminated). The oversampling

stage relies on the spheres produced during the cleaning stage. For every minority example, new examples are generated within its sphere, where the proportion of examples to generate is inversely proportional to the radius of its sphere. **G-SMOTE** replaces the interpolation method used by SMOTE to define a flexible geometric region (a truncated hyperspheroid) where the synthetisation of new examples occurs [37]. A minority example and one of its closest nearest minority neighbours are used to define a unit hypersphere where the new synthetic example will be generated. Through a set of geometric hyperparameters, the hypersphere can be transformed to represent different configurations (hyperspheroids) and parameters can be tuned for optimal performance.

Overall, we are witnessing a shift towards approaches that combine multiple sources of information (local and structural information) and couple different operations to achieve optimal results. The main objective is that new approaches address the existing limitations of their predecessors, while increasingly adapting to the characteristics of the domains.

7.4 Other Approaches

Undersampling, oversampling, and cleaning approaches are by far the most common in the field. Herein, we discuss other emergent approaches to handle imbalanced and overlapped domains. These are based on different paradigms, namely **Ensembles, Region Splitting, Evolutionary and Hybrid Approaches**.

Ensembles are based on the combination of different classifiers, called *base classifiers*. Each base classifier is trained over the data domain and the individual predictions are combined to produce the final decision. The model resulting from that aggregation is the *ensemble*, which is then used to classify new data examples [146]. In imbalanced learning, popular ensembles are Boosting (commonly AdaBoost) [49, 50] and Bagging [15]. However, the traditional use of ensembles (simply combining classifiers) hardly solves the class imbalance problem by itself, let alone handle both imbalanced and overlapped domains [44]. On contrary, ensembles are commonly coupled with resampling (undersampling or oversampling), and cleaning strategies, in order to adapt to the peculiarities of the domains.

Chen et al. [23] produce a software defect prediction model **(SDPM)** that combines class overlap reduction and ensemble imbalance learning. First, NCL cleaning is used to remove the overlapping examples. Then, the data is randomly undersampled several times to produce different subsets that are trained by different classifiers. The final classification model is built by assembling the base classifiers through the AdaBoost mechanism. **CluAD-EdiDO** [26] was developed to handle multi-class imbalanced and overlapped datasets. First, a clustering-based adaptive decomposition is applied to generate an adaptive number of clusters. Then, an editing-based diversified oversampling method is used to address class imbalance and overlapping in different clusters. For the overlapping problem, a cleaning technique is used (removing examples with complicated neighbourhoods) whereas the class imbalance problem is alleviated by SMOTE or DKNOS [26, 112], depending on the type of example. Finally, an ensemble learning framework is used to select the best classification algorithm for each cluster.

Region Splitting approaches (same as *separating scheme* approaches [147]) separate the data domain into non-overlapping and overlapping regions (or safe and overlapping regions). Then, each region is handled independently, by different classifiers or using different parametrisations of the same classifier (e.g., different k values in kNN, different SVM hyperparameters) [127, 128].

In the last couple of years, this "divide-and-conquer" strategy has been popular in imbalanced and overlapped domains. **Soft-Hybrid** [131] divides the data domain into non-overlapped, borderline, and overlapped regions using the modified Hausdorff distance [67], Radial Basis Function Networks (RBFN), and k-means. After the boundaries of each region are found, DBSCAN is applied to the borderline regions, whereas RBFNs are considered for the remaining. **OSM** [81] separates the data space into soft and hard overlap regions. Soft-overlap regions are classified using the decision boundary of the OSM classifier (a modified fuzzy SVM), whereas hard-overlap regions are classified using 1NN. An important feature of OSM is the integration of instance-level overlap (defined using kNN) and global information regarding class imbalance (via the Different Error Cost algorithm [12]) to produce overlap-sensitive costs (weights) that are further incorporated in its optimisation function.

Evolutionary Algorithms (EAs) are nature-inspired solutions, often associated to biological processes, such as reproduction, mutation, and recombination [119]. The process of finding an optimal solution is based on a natural selection mechanism: the weakest solutions are eliminated whereas the strongest are retained in the subsequent evolutions. In imbalanced and overlapped domains, EAs are used to select a representative set of examples from the training set that simultaneously minimise the imbalance ratio, improve the representation of the minority class in overlap regions, and avoid information loss.

EVINCI [42] uses a multi-objective evolutionary algorithm (NSGA-II [35]) to selectively reduce the concentration of redundant majority examples in the overlapping areas, thus improving the representation of minority examples in these areas. **EHSO** [153] finds overlapping regions by analysing the local neighbourhood of each majority example. If a given majority example has at least one minority class neighbour, then it is considered an overlapping example. Then, overlapping majority examples are removed in a way that the decision boundary between classes is maximised while preserving the original data information as much as possible through the use of CHC evolutionary algorithm [39].

Finally, **Hybrid** approaches may aggregate a series of features from the previous methods. As discussed throughout this section, several of the listed approaches can be considered an hybridisation of others. For instance, certain oversampling approaches have a data cleaning component (e.g., SMOTE-TL), while ensembles, region splitting and EA approaches are often coupled with resampling (oversampling, undersampling), and cleaning techniques. Herein, we highlight recent hybrid approaches explored in the context of imbalanced and overlapped domains.

MBP-GGE [3] uses a modified back-propagation multilayer perceptron to improve the visibility of the minority class during the training process. Additionally, it eliminates majority examples in overlapping regions using a the Gabriel Graph Editing technique (GGE). **BoostOBU** improves the detection of majority class examples in the overlapping region, reducing excessive elimination [132]. First, it applies Borderline-SMOTE to emphasize the minority class borders. Then,

AdaOBU is applied. **ImWeights** [79] combines structural and local information to preprocess imbalanced data by simultaneous clustering and categorising minority examples. First, it uses ImGrid clustering [78] to produce a grid of cells containing information on the types of minority examples and existing minority sub-clusters. Then, examples are weighted according to both their safety and their distance to neighbouring minority clusters, using a *gravity* concept. The final weights can then be incorporated into the learning process of classifiers.

7.5 Summarising Comments

Throughout the previous sections, we have carried out a thorough review of the state-of-the-art class overlap-based approaches used in imbalanced domains. Additionally, we proposed a new taxonomy of methods that resonates with the representations of class overlap they are associated to. Overall, it is possible to identify some trends regarding class overlap-based methods, which we summarise in what follows.

Undersampling approaches are more prone to consider structural information, via clustering and graph-based approaches. These strategies are used to establish the regions of interest of the data domains (core concepts) and discard redundant and overlapped examples.

Alternatively, cleaning and oversampling approaches prioritise local information, mostly evaluating instance-level overlap. In cleaning approaches, the value of k determines the depth of the cleaning procedure (either addressing borderline regions or the entire domain). To this regard, multiresolution information (fine-grain search) has also been explored successfully to recursively remove harmful examples.

Oversampling is increasingly moving towards parametrised approaches that adapt the generation of new examples to the characteristics of data. There is also some concern with the generation of examples that are both informative and diverse (e.g., PAIO, G-SMOTE). 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 parameters, for which there is not yet an established relationship with data characteristics.

Finally, is not uncommon for approaches to share some paradigms (e.g., local, structural, and density information, fuzzy logic, and cost-sensitive strategies). This goes towards the idea that class overlap has different vortices of complexity, and addressing them altogether could potentially improve results. Also, there is a considerably lower number of approaches developed within the scope of ensembles, evolutionary, region splitting, and hybrid approaches, which may be due to the lack of current knowledge on the joint-effect of class imbalance and overlap on different learning paradigms. This motivates the need to put forward some insights regarding the footprints of different families of classifiers, as we have performed in Section 4.

Nevertheless, as stated at the introduction of this section, it is still premature to derive recommendations for researchers regarding class overlap-based methods on the basis of related research. On that note, Table 2 provides an overview of

class overlap-based approaches, referring to the proposed taxonomy, the information considered by each approach, the type of data characterisation provided (i.e., whether both class imbalance and overlap are measured and how), and the benchmark of methods used for comparison.

Table 2: Benchmark of class overlap-based approaches. For each approach is identified its category, the type of information it encompasses, the considered measures of class imbalance and class overlap, and a benchmark of compared methods. Approaches are marked depending on whether they obtained superior performance with respect to F-measure/G-mean results (in bold), sensitivity results (†) or AUC results (‡).

Category	Approach	Information	Measures	Compared Methods
Undersampling	**ClusBUS**†‡ (2014)	Density-based clustering	IR	SMOTE
	DBMUTE‡ (2017)	Density-based clustering Graph-based	IR	ROS, RUS, SMOTE, BLSMOTE SLSMOTE, DBSMOTE, TL MUTE
	DBMIST-US (2020)	Density-based clustering Graph-based	IR	CNN, ENN, TL, NCL, OSS SBC, ClusterOSS, RUS, EUS EE, BC, RUSBoost
	ClusterOSS (2014)	Cluster-based (k-means) Local Information (1NN)	IR	OSS, RUS†, ROS, SMOTE, CBO **ClusterOSS+ROS**‡
	CUST‡ (2016)	Cluster-based (k-means) Local Information (1NN)	IR	RUS, ROS, ClusBUS, SMOTE, OSS
	OBU† (2018)	Fuzzy-based clustering	IR	kmUnder
	AdaOBU† (2020)	Fuzzy-based clustering Adaptive threshold	IR	SMOTE, BLSMOTE, kmUnder SMOTE-ENN, **SMOTEBag** RUSBoost, OBU, BoostOBU
Cleaning	**MUTE** (2011)	Local Information (kNN)	IR	BLSMOTE, SLSMOTE, SMOTE
	SMOTE-IPF‡ (2015)	Local Information (kNN) Ensemble-based Fine-Grain Search	IR	SMOTE, SMOTE-TL, SMOTE-ENN SLSMOTE, BLSMOTE
	NB-Basic (2020) NB-Tomek (2020) **NB-Comm** (2020) NB-Rec† (2020)	Local Information (1NN) Local Information (kNN) Local Information (kNN) Local Information (kNN) Fine-Grain Search	IR	**SMOTE**, BLSMOTE, ENN kmUnder, OBU
Oversampling	**MWMOTE**‡ (2014)	Cluster-based (hierarchical) Density information Local information (kNN)	IR	SMOTE, ADASYN, RAMOBoost
	ASUWO‡ (2016)	Cluster-based (hierarchical) Local information (kNN) Classification Complexity	IR	ROS, SMOTE, BLSMOTE, SLSMOTE kmUnder, ClusterSMOTE, CBO MWMOTE
	IA-SUWO‡ (2020)	Cluster-based (hierarchical) Local information (kNN) Classification Complexity Adaptive Weighting	IR	ROS, SMOTE, BLSMOTE, ADASYN SLSMOTE, ClusterSMOTE, MWMOTE A-SUWO, ISMOTE, kmSMOTE
	NI-MWMOTE†‡ (2020)	Cluster-based (hierarchical) Local information (kNN) Classification Complexity Density information	IR	ROS, SMOTE, BLSMOTE, ADASYN SLSMOTE, ClusterSMOTE MWMOTE, A-SUWO
	PAIO†‡ (2020)	Density-based clustering Local information (kNN)	IR	ROS, SPIDER, SMOTE, SLSMOTE MWMOTE, SMOM, INOS, MDO RACOG
	CCR†‡ (2017)	Hypersphere Coverage	IR	SMOTE, ADASYN, BLSMOTE SMOTE-TL, SMOTE-ENN, NCL
	G-SMOTE‡ (2019)	Hypersphere Coverage	IR	ROS, SMOTE
	SDPM†‡ (2018)	Ensemble-based Local Information (kNN) Undersampling	IR	EE, NBLog, RF, NB, SMOTE+NB RUS+NB, DNC, SMOTEBoost RUSBoost
	CluAD-EdiDO (2020)	Ensemble-based Cluster-based Local information (kNN) Oversampling	IR and OR	SMOTE, **SMOTEBag**, RUS, ROS RUSBoost, KNOS, DOVO, DOAO MDO, DECOC, GP-ECOC
	Soft-Hybrid† (2015)	Region Splitting Cluster-based Local and Density information	IR and F1	SVM, RBFN SVM/RBFN:(ROS, RUS, SMOTE)
	OSM (2018)	Region Splitting Fuzzy Logic (Fuzzy SVM) Cost-sensitive Local Information (kNN and 1NN)	IR and OR	SVM, SVM+RUS, SMOTE-SVM, SDC SVMBoost, FSVM-CIL, EFSVM EMatMHKS, 1NN

To be continued on the next page...

Table 2: Continued from previous page.

Category	Approach	Information	Measures	Compared Methods
Other Approaches	**EVINCI** (2019)	Evolutionary-based Ensemble-based Graph-based Local Information (1NN)	IR and N1	SMOTEBag, RUSBag, ROSBag Adaboost, RUSBoost
	EHSO‡ (2020)	Evolutionary-based Local Information (kNN) Undersampling	IR and OR	RUS, NCL, NM, IHT, RENN, AKNN OSS, ROS, SMOTE, BLSMOTE ADASYN, SMOTE-ENN, SMOTE-TL RBO, SMOTE-CCA, CCR
	MBP-GGE (2013)	Hybrid Approach Graph-based Cost-sensitive	IR	SBP, MBP, SBP+GGE, **SMOTE**, RUS SMOTE+GGE
	BoostOBU (2020)	Hybrid Approach Fuzzy-based clustering Local Information (kNN) Oversampling Undersampling	IR	SMOTE, BLSMOTE, kmUnder SMOTE-ENN, **SMOTEBag**, RUSBoost OBU, AdaOBU†
	ImWeights (2018)	Hybrid Approach Cluster-based Local information (kNN) Cost-sensitive	IR and Data Typology	ROS, BLSMOTE, ADASYN

†: The approach obtained superior performance with respect to sensitivity results.
‡: The approach obtained superior performance with respect to AUC results.
OR refers to Overlapping Ratio, which may differ between approaches (please refer to the discussion).
EUS[54], EE[85], BC[85], RUSBoost[113], kmUnder[149], SMOTEBag[139], RAMOBoost[25], Cluster-SMOTE[27], ISMOTE[24] kmSMOTE[38], INOS[20], MDO[1], SMOM[151], RACOG[33], NBLog [97], DNC[140], SMOTEBoost[22], SDC[2] SVMBoost[138], FSVM-CIL[12], EFSVM[41], EMatMHKS[150], RUSBag[6], ROSBag[139], NM[93], IHT[120], RENN[77] AKNN[77], RBO[75], SMOTE-CCA[148], KNOS[112], DOVO[52], DOAO[71], DECOC[13], GP-ECOC[83].

Let us first discern why it is not possible to support the application of one approach (or category of approaches) over the others from a theoretical point of view, i.e., based on the internal behaviour of approaches. First, 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. This may be been attenuated to some extent by the development of more refined approaches, but at the cost of increasing computational complexity and interpretability (too many user-defined parameter to tune). Secondly, the advantage of oversampling techniques due to their ability of considering the inner structure of data [59] may not hold for imbalanced and overlapped domains. Indeed, most recent undersampling and cleaning approaches also comprise structural and local information of the domains and have proven to surpass well-established oversampling algorithms (Table 2). Finally, there are obvious advantages in using other types of approaches, such as the incorporation of data complexity and classification performance in multi-objective evolutionary approaches, or the combination of multiple reasoning paradigms when using ensembles.

There are further limitations found in current research that make it impossible to provide an evidence-based recommendation of strategies to handle imbalanced and overlapped domains. Let us conclude this section by discussing the most important.

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 with their analogous class imbalance/distribution-based approaches, rather than approaches developed for the same purpose (i.e., handling both class imbalance and overlap). Thus, it would be informative to compare approaches of the same category (e.g., DBMIST-US versus AdaOBU), as well as approaches of different categories (e.g., DBMIST-US, NB-Comm, and NI-MWMOTE).

Furthermore, despite many methods are being proposed to overcome class overlap, there is a clear lack of information on how datasets are affected by this problem, i.e., only a few works provide a characterisation of class overlap. In fact, in most of the related work, the used datasets are not characterised beyond their number of examples, features, and imbalance ratio (Table 2). In terms of improvements with respect to class overlap, the approaches are evaluated from a theoretical perspective, according to their inner behaviour and the effects of their application on classification performance, and without real empirical validation. It is suggested that class overlap is alleviated since the classification results improve, although no class overlap measures are analysed to support such claim. Hence, it would be crucial to evaluate class overlap measures before and after the application of methods to fully characterise their ability to solve the problem and perform a fair comparison between approaches.

Finally, since no standard measure of class overlap is yet established, related research resorts to different measures to characterise the domains, similarly to what was observed for seminal work on synthetic datasets (Section 2). Some works refer to specific measures (F1, N1, or data typology), while others refer to a generic Overlapping Ratio (OR), which is based on different variations of instance-level overlap measures. Beyond not using a standard measurement of class overlap, related work is in fact focusing on distinct vortices of class overlap, by using measures that capture different dimensions of the problem. Again, it becomes clear that there is much to be explored regarding the joint-effect of class imbalance and overlap, and why a unified view on the problem is necessary for perceptive advances in the field.

8 Open Challenges

Class overlap is currently one of the major difficulty factors affecting classification performance in imbalance domains. Although previous research was able to establish some insights regarding the joint-impact of class imbalance and overlap on classification performance, the critical analysis presented in this work shows that there is still a lot to uncover. As discussed throughout Section 5, seminal work on synthetic data suffered from three major shortcomings, which have not yet been completely solved for real-world domains (as discussed in Sections 6 and 7):

Class overlap is not mathematically well-established:
Contrary to class imbalance, there is not a well-established formulation and measurement of class overlap for real-world domains, despite the fact that several data complexity measures have been discussed throughout the years. This leads to the lack of characterisation of class overlap across recent research and prevents a deeper analysis and comparison of proposed approaches.

Class overlap assumes different representations:
Due to the lack of a standard measurement of class overlap, related research on real-world domains uses different measures that may be focusing on distinct vortices of the problem, which further complicates the comparison between approaches. Nevertheless, it is possible to associate the underlying principles of existing class overlap-based approaches to the class overlap representations they are sensitive to. Thoroughly characterising class overlap in real-world

domains would be instrumental to guide the choice of appropriate approaches and the development of specialised methods.

The class overlap degree does not take other factors into account: Recent advances in the field show that there is an increasing interest in the study of class overlap measures that account for other characteristics of data, especially class imbalance [8, 9]. Some well-established measures have recently proved to be biased indicators in the presence of class imbalance, and consequently new adaptations are starting to emerge. Beyond class imbalance, it seems that future research will gravitate more and more around the idea that class overlap comprises multiple sources of complexity, and that new measures need to account for its heterogeneous nature [105].

In this work, we provide a comprehensive and unique view on the joint-effect of class imbalance and overlap, and discuss new perspectives in light of the limitations found in related work. In sum, the research community needs to move towards a unified view of the problem of class overlap in imbalanced domains regarding three main topics:

1. **Representations of class overlap:**
   It is important that the research community comes together in establishing important concepts associated with class overlap and defining the types of degradation they are associated to, i.e., their impact on classification performance. To this regard, the ideas explored in this work regarding distinct representations of class overlap aim to start the discussion among researchers. Following directions should be taken in order to fully understand the problem of class overlap in real-world domains:

   - The study of public repositories (e.g., UCI[5], Kaggle[6], KEEL[7], OpenML[8]) in what concerns the analysis of data intrinsic characteristics would be an important contribution to future research. With respect to the problem of class overlap, the taxonomy provided in Section 6 allows to group datasets depending on their dominant overlap representation. Accordingly, some domains may be conceptually intertwined (structural overlap), whereas others may be mostly affected by complicated examples (referring to instance-level overlap). 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. Also with respect to the established representations of class overlap, it would be interesting to study the effect of each type of degradation (and their combination) on the performance of classifiers with distinct learning paradigms.
   - The enhancement of existing repositories with artificial datasets (or modification of real-world datasets via data morphing [29, 109] or evolutionary algorithms [48, 91, 96, 99]) is also a possibility for future research. In such

[5] `https://archive.ics.uci.edu`
[6] `https://www.kaggle.com`
[7] `http://keel.es`
[8] `https://www.openml.org`

a way, the diversity of current repositories can be improved by tailoring the new datasets to specific sources and ranges of data complexity (e.g., introducing specific vortices of class overlap, more complex data structures, and class skews).

- In the scope of artificial data generation, we recommend the multidimensional data generator described in [145], 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.[9] We welcome other researchers to contribute with their own research data in order to move towards the creation of a representative repository of data complexity factors, beyond imbalanced and overlapped datasets.

2. **Characterisation and quantification of class overlap:**
Future research should keep moving towards the definition of measures with broader points of view, i.e., that are able to combine different representations of class overlap and consider other factors, mainly class imbalance. On that note, the discussion presented in Section 6 can serve as stepping stone. It provides an overview of existing class overlap measures and the class overlap representations they are associated to, the type of insights they provide, and whether they consider additional complications (e.g., class imbalance). The following directions may guide future researchers towards a better insight into the characterisation of the class overlap problem in imbalanced domains:

   - Acknowledging class overlap as a heterogeneous concept, the development of new measures that combine several sources of complexity/information is perhaps the most pressing topic for future research. To this point, existing complexity measures focus on assessing individual properties of data, whereas real-world domains require more perceptive and flexible sets of measures. In that regard, our proposed taxonomy may be a starting point to the exploration of measures with broader points of view, namely in what concerns the combination of class overlap representations and associated insights.
   - Beyond the measures identified in Figure 4 and highlighted in Section 6.5, which have been designed or adapted to account for class imbalance, the remaining should be further investigated in imbalanced domains.
   - The development of approaches to assess other learning tasks other than binary-classification problems, namely multi-class domains, also remains a topic for future research. Most class overlap measures are studied over binary-classification domains, and current adaptations to class imbalance (i.e., class decomposition) may not be adequate to the evaluation of multi-class problems [26, 53, 103, 109].

3. **Benchmark of approaches for imbalanced and overlapped domains:**
It would be important to provide a benchmark of approaches that simultaneously handle class imbalance and overlap, in light of the ideas discussed throughout the paper. It is crucial to compare state-of-the-art approaches with

[9] `https://github.com/miriamspsantos/datagenerator`

each other, rather than with well-established methods. Also, a more insightful characterisation of datasets is necessary. It is fundamental to fully characterise the problem of class overlap in the domains, so that improvements introduced by the approaches are more profoundly assessed. Also, the characterisation of domains is essential to infer on the behaviour of approaches with distinct underlying mechanisms. To this regard, the summary of existing benchmarks and the taxonomy proposed in Section 7 is a good starting direction. The development of new approaches for handling imbalanced and overlapped domains may take into consideration the following directions:

- Future research should evaluate new proposed approaches against emergent methods developed during recent years, rather than limiting the analysis to well-established approaches. It is also important to consider a deeper characterisation of datasets, beyond the number of examples, features, and imbalance ratio. The same is true regarding the standardisation of performance metrics. These aspects are crucial to guarantee a fair evaluation and comparison of approaches.
- A large number of class overlap-based approaches is based on the evaluation of complicated examples (e.g., borderline, noisy examples), mostly relying on the assessment of instance-level overlap. New studies in the field should explore other vortices of class overlap simultaneously, to produce more robust solutions.
- Future work should consider sharing the source code and obtained results of proposed approaches, in order to guarantee the reproducibility of research results. Regarding imbalanced and overlapped domains, we provide a collection of related resources (data and code), which researchers may consider in future experiments.[10] Additionally, we provide an extended Python library – *Python Class Overlap Library* (`pycol`)[11] – comprising the class overlap complexity measures discussed in Section 6, to encourage a more comprehensive study of the problem of class overlap.

Addressing these avenues would provide a renewed and improved view on the problem, ultimately leading to important advances in the field.

9 Conclusions

In this work, we address the joint-effect of class imbalance and overlap in classification tasks, from precursor work to most emergent approaches, showing that their combination is still not completely understood. Accordingly, the paper may be divided into two main parts.

First, we start by discussing the insights derived from previous work on the topic, as well as existing limitations. We focus particularly on the analysis of some neglected, although important, aspects left undiscussed in seminal research, namely *i)* the influence of intrinsic data characteristics (data decomposition, data structure, data dimensionality, data typology) on the classification performance for imbalanced and overlapped domains, and *ii)* the characterisation of the footprints

[10] `https://github.com/miriamspsantos/open-source-imbalance-overlap`
[11] `https://github.com/miriamspsantos/pycol`

of classifiers with distinct learning biases in this context. The analysis of related research culminated in the identification of limitations regarding the characterisation of the problem of class overlap in real-world domains and finally, to the acknowledgement of class overlap as a heterogeneous concept, comprising multiple sources of complexity.

Accordingly, we move towards the second part of this work, discussing the key concepts associated to the identifiability and quantification of class overlap, and the most recent approaches to address the problem in real-world domains. In that regard, we first propose a novel taxonomy of class overlap complexity measures, comprising four main class overlap representations: Feature Overlap, Structural Overlap, Instance-Level Overlap, and Multiresolution Overlap. A comprehensive set of complexity measures associated with class overlap is thoroughly reviewed, and each measure is included in one of the established groups, depending on which representation it is able to capture. Then, the most emergent class overlap-based approaches in imbalanced domains are analysed following the same perspective: we further present a taxonomy of class overlap-based approaches associating their underlying behaviour to the class overlap representations they are attentive to. In other words, the taxonomy of class overlap-based approaches is aligned with the established taxonomy of class overlap complexity measures.

In sum, this work provides a global and unique view on the joint-problem of class imbalance and overlap, discussing important concepts from related research, exploring new perspectives in light of the limitations found, and establishing key insights that may hopefully encourage future researchers to move towards a unified view on the problem and inspire the development of novel approaches that account for the peculiarities of imbalanced and overlapped domains.

Acknowledgements This work is funded by national funds through the FCT - Foundation for Science and Technology, I.P., within the scope of the project CISUC - UID/CEC/00326/2020 and by European Social Fund, through the Regional Operational Program Centro 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. This work was also partially funded by the project Safe Cities - Inovação para Construir Cidades Seguras, with the reference POCI-01-0247-FEDER-041435, co-funded by the European Regional Development Fund (ERDF), through the Operational Programme for Competitiveness and Internationalization (COMPETE 2020), under the PORTUGAL 2020 Partnership Agreement. The work is further supported by the FCT Research Grant SFRH/BD/138749/2018.

Declarations

Funding
Not applicable.
Conflicts of interest/Competing interests
The authors declare that they have no conflict of interest.
Availability of data and material
Not applicable.
Code availability
`https://github.com/miriamspsantos/pycol`
Authors' contributions
Miriam Seoane Santos: Conceptualisation, Methodology, Literature Search,

Investigation, Formal Analysis, Writing - Original Draft, Writing - Review and Editing, Visualisation. **Pedro Henriques Abreu:** Conceptualisation, Validation, Writing - Review and Editing, Supervision. **Nathalie Japkowicz:** Validation, Writing - Review and Editing. **Alberto Fernández:** Validation, Writing - Review and Editing. **Carlos Soares:** Validation, Writing - Review and Editing. **Szymon Wilk:** Validation, Writing - Review and Editing. **João Santos:** Writing - Review and Editing.

References

1. Abdi L, Hashemi S (2015) To combat multi-class imbalanced problems by means of over-sampling techniques. IEEE transactions on Knowledge and Data Engineering 28(1):238–251
2. Akbani R, Kwek S, Japkowicz N (2004) Applying support vector machines to imbalanced datasets. In: European conference on machine learning, Springer, pp 39–50
3. Alejo R, Valdovinos RM, García V, Pacheco-Sanchez JH (2013) A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios. Pattern Recognition Letters 34(4):380–388
4. Anwar N, Jones G, Ganesh S (2014) Measurement of data complexity for classification problems with unbalanced data. Statistical Analysis and Data Mining: The ASA Data Science Journal 7(3):194–211
5. Armano G, Tamponi E (2016) Experimenting multiresolution analysis for identifying regions of different classification complexity. Pattern Analysis and Applications 19(1):129–137
6. Barandela R, Valdovinos RM, Sánchez JS (2003) New applications of ensembles of classifiers. Pattern Analysis & Applications 6(3):245–256
7. Barella VH, Costa EP, Carvalho A, PL F (2014) Clusteross: a new undersampling method for imbalanced learning. In: Proc. of 3th Brazilian Conference on Intelligent Systems. Academic Press
8. Barella VH, Garcia LP, de Souto MP, Lorena AC, de Carvalho A (2018) Data complexity measures for imbalanced classification tasks. In: 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, pp 1–8
9. Barella VH, Garcia LP, de Souto MC, Lorena AC, de Carvalho AC (2021) Assessing the data complexity of imbalanced datasets. Information Sciences 553:83–109
10. Barua S, Islam M, Yao X, Murase K (2014) Mwmote–majority weighted minority oversampling technique for imbalanced data set learning. IEEE Transactions on Knowledge and Data Engineering 26(2):405–425
11. Batista GE, Prati RC, Monard MC (2004) A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter 6(1):20–29
12. Batuwita R, Palade V (2010) Fsvm-cil: fuzzy support vector machines for class imbalance learning. IEEE Transactions on Fuzzy Systems 18(3):558–571
13. Bi J, Zhang C (2018) An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme. Knowledge-Based Systems 158:81–93

14. Borsos Z, Lemnaru C, Potolea R (2018) Dealing with overlap and imbalance: a new metric and approach. Pattern Analysis and Applications 21(2):381–395
15. Breiman L (1996) Bagging predictors. Machine learning 24(2):123–140
16. Bunkhumpornpat C, Sinapiromsaran K (2017) Dbmute: density-based majority under-sampling technique. Knowledge and Information Systems 50(3):827–850
17. Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C (2009) Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In: Pacific-Asia conference on knowledge discovery and data mining, Springer, pp 475–482
18. Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C (2011) Mute: Majority under-sampling technique. In: 2011 8th International Conference on Information, Communications & Signal Processing, IEEE, pp 1–4
19. Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C (2012) Dbsmote: density-based synthetic minority over-sampling technique. Applied Intelligence 36(3):664–684
20. Cao H, Li XL, Woon DYK, Ng SK (2013) Integrated oversampling for imbalanced time series classification. IEEE Transactions on Knowledge and Data Engineering 25(12):2809–2822
21. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research 16:321–357
22. Chawla NV, Lazarevic A, Hall LO, Bowyer KW (2003) Smoteboost: Improving prediction of the minority class in boosting. In: European conference on principles of data mining and knowledge discovery, Springer, pp 107–119
23. Chen L, Fang B, Shang Z, Tang Y (2018) Tackling class overlap and imbalance problems in software defect prediction. Software Quality Journal 26(1):97–125
24. Chen S (2017) An improved synthetic minority over-sampling technique for imbalanced data set learning. Degree Thesis of Department of Information Engineering, National Tsing Hua University pp 1–59
25. Chen S, He H, Garcia EA (2010) Ramoboost: Ranked minority oversampling in boosting. IEEE Transactions on Neural Networks 21(10):1624–1642
26. Chen X, Zhang L, Wei X, Lu X (2020) An effective method using clustering-based adaptive decomposition and editing-based diversified oversamping for multi-class imbalanced datasets. Applied Intelligence pp 1–16
27. Cieslak DA, Chawla NV, Striegel A (2006) Combating imbalance in network intrusion datasets. In: GrC, Citeseer, pp 732–737
28. Cohen G, Hilario M, Sax H, Hugonnet S, Geissbuhler A (2006) Learning from imbalanced data in surveillance of nosocomial infection. Artificial intelligence in medicine 37(1):7–18
29. Correia A, Soares C, Jorge A (2019) Dataset morphing to analyze the performance of collaborative filtering. In: International Conference on Discovery Science, Springer, pp 29–39
30. Costa AJ, Santos MS, Soares C, Abreu PH (2020) Analysis of imbalance strategies recommendation using a meta-learning approach. In: 7th ICML Workshop on Automated Machine Learning (AutoML-ICML2020), pp 1–10
31. Cummins L (2013) Combining and choosing case base maintenance algorithms. PhD thesis, University College Cork

32. Das B, Krishnan NC, Cook DJ (2014) Handling imbalanced and overlapping classes in smart environments prompting dataset. In: Data mining for service, Springer, pp 199–219
33. Das B, Krishnan NC, Cook DJ (2014) Racog and wracog: Two probabilistic oversampling techniques. IEEE transactions on knowledge and data engineering 27(1):222–234
34. Das S, Datta S, Chaudhuri B (2018) Handling data irregularities in classification: Foundations, trends, and future challenges. Pattern Recognition 81:674–693
35. Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast and elitist multiobjective genetic algorithm: Nsga-ii. IEEE transactions on evolutionary computation 6(2):182–197
36. Denil M, Trappenberg T (2010) Overlap versus imbalance. In: Canadian Conference on Artificial Intelligence, Springer, pp 220–231
37. Douzas G, Bacao F (2019) Geometric smote a geometrically enhanced drop-in replacement for smote. Information sciences 501:118–135
38. Douzas G, Bacao F, Last F (2018) Improving imbalanced learning through a heuristic oversampling method based on k-means and smote. Information Sciences 465:1–20
39. Eshelman LJ (1991) The chc adaptive search algorithm: How to have safe search when engaging in nontraditional genetic recombination. In: Foundations of genetic algorithms, vol 1, Elsevier, pp 265–283
40. Ester M, Kriegel HP, Sander J, Xu X, et al. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd, vol 96, pp 226–231
41. Fan Q, Wang Z, Li D, Gao D, Zha H (2017) Entropy-based fuzzy support vector machine for imbalanced datasets. Knowledge-Based Systems 115:87–99
42. Fernandes ER, de Carvalho AC (2019) Evolutionary inversion of class distribution in overlapping areas for multi-class imbalanced learning. Information Sciences 494:141–154
43. Fernández A, García S, Galar M, Prati R, Krawczyk B, Herrera F (2018) Data Intrinsic Characteristics, Springer International Publishing, Cham, pp 253–277
44. Fernández A, García S, Galar M, Prati R, Krawczyk B, Herrera F (2018) Ensemble Learning, Springer International Publishing, Cham, pp 147–196
45. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F (2018) Dimensionality reduction for imbalanced learning. In: Learning from Imbalanced Data Sets, Springer, pp 227–251
46. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F (2018) Learning from imbalanced data sets, vol 11. Springer
47. Fernández A, Garcia S, Herrera F, Chawla NV (2018) Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of artificial intelligence research 61:863–905
48. França TR, Miranda PB, Prudêncio RB, Lorenaz AC, Nascimento AC (2020) A many-objective optimization approach for complexity-based data set generation. In: 2020 IEEE Congress on Evolutionary Computation (CEC), IEEE, pp 1–8

49. Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences 55(1):119–139
50. Friedman J, Hastie T, Tibshirani R, et al. (2000) Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Annals of statistics 28(2):337–407
51. Fu GH, Wu YJ, Zong MJ, Yi LZ (2020) Feature selection and classification by minimizing overlap degree for class-imbalanced data in metabolomics. Chemometrics and Intelligent Laboratory Systems 196:103906
52. Galar M, Fernández A, Barrenechea E, Bustince H, Herrera F (2013) Dynamic classifier selection for one-vs-one strategy: avoiding non-competent classifiers. Pattern Recognition 46(12):3412–3424
53. Galar M, Fernández A, Barrenechea E, Herrera F (2015) Drcw-ovo: distance-based relative competence weighting combination for one-vs-one strategy in multi-class problems. Pattern recognition 48(1):28–42
54. García S, Herrera F (2009) Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. Evolutionary computation 17(3):275–306
55. García V, Alejo R, Sánchez J, Sotoca J, Mollineda R (2006) Combined effects of class imbalance and class overlap on instance-based classification. In: International Conference on Intelligent Data Engineering and Automated Learning, Springer, pp 371–378
56. García V, Mollineda R, Sánchez J, Alejo R, Sotoca J (2007) When overlapping unexpectedly alters the class imbalance effects. In: Iberian Conference on Pattern Recognition and Image Analysis, Springer, pp 499–506
57. García V, Sánchez J, Mollineda R (2007) An empirical study of the behavior of classifiers on imbalanced and overlapped data sets. In: Iberoamerican Congress on Pattern Recognition, Springer, pp 397–406
58. García V, Mollineda R, Sánchez J (2008) On the k-nn performance in a challenging scenario of imbalance and overlapping. Pattern Analysis and Applications 11(3-4):269–280
59. García V, Sánchez J, Marqués A, Florencia R, Rivera G (2020) Understanding the apparent superiority of over-sampling through an analysis of local information for class-imbalanced data. Expert Systems with Applications 158:113026
60. Greene J (2001) Feature subset selection using thornton's separability index and its applicability to a number of sparse proximity-based classifiers. In: Proceedings of Annual Symposium of the Pattern Recognition Association of South Africa
61. Guzmán-Ponce A, Valdovinos RM, Sánchez JS, Marcial-Romero JR (2020) A new under-sampling method to face class overlap and imbalance. Applied Sciences 10(15):5164
62. Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G (2017) Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications 73:220–239
63. Han H, Wang WY, Mao BH (2005) Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: International conference on intelligent computing, Springer, pp 878–887

64. Hart P (1968) The condensed nearest neighbor rule (corresp.). IEEE transactions on information theory 14(3):515–516
65. He H, Bai Y, Garcia E, Li S (2008) Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on, IEEE, pp 1322–1328
66. Ho T, Basu M (2002) Complexity measures of supervised classification problems. IEEE transactions on pattern analysis and machine intelligence 24(3):289–300
67. Huttenlocher DP, Klanderman GA, Rucklidge WJ (1993) Comparing images using the hausdorff distance. IEEE Transactions on pattern analysis and machine intelligence 15(9):850–863
68. Jain A, Duin R, Mao J (2000) Statistical pattern recognition: A review. IEEE Transactions on pattern analysis and machine intelligence 22(1):4–37
69. Japkowicz N (2001) Concept-learning in the presence of between-class and within-class imbalances. In: Conference of the Canadian society for computational studies of intelligence, Springer, pp 67–77
70. Jo T, Japkowicz N (2004) Class imbalances versus small disjuncts. ACM Sigkdd Explorations Newsletter 6(1):40–49
71. Kang S, Cho S, Kang P (2015) Constructing a multi-class classifier using one-against-one approach with different binary classifiers. Neurocomputing 149:677–682
72. Kaur H, Pannu HS, Malhi AK (2019) A systematic review on imbalanced data challenges in machine learning: Applications and solutions. ACM Computing Surveys (CSUR) 52(4):1–36
73. Kovács G (2019) An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. Applied Soft Computing 83:105662
74. Koziarski M, Wozniak M (2017) Ccr: A combined cleaning and resampling algorithm for imbalanced data classification. International Journal of Applied Mathematics and Computer Science 27(4):727–736
75. Koziarski M, Krawczyk B, Wozniak M (2019) Radial-based oversampling for noisy imbalanced data classification. Neurocomputing 343:19–33
76. Krawczyk B (2016) Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence 5(4):221–232
77. Kubat M, Matwin S, et al. (1997) Addressing the curse of imbalanced training sets: one-sided selection. In: Icml, Citeseer, vol 97, pp 179–186
78. Lango M, Brzezinski D, Firlik S, Stefanowski J (2017) Discovering minority sub-clusters and local difficulty factors from imbalanced data. In: International Conference on Discovery Science, Springer, pp 324–339
79. Lango M, Brzezinski D, Stefanowski J (2018) Imweights: Classifying imbalanced data using local and neighborhood information. In: Second International Workshop on Learning with Imbalanced Domains: Theory and Applications, PMLR, pp 95–109
80. Laurikkala J (2001) Improving identification of difficult small classes by balancing class distribution. In: Conference on Artificial Intelligence in Medicine in Europe, Springer, pp 63–66
81. Lee HK, Kim SB (2018) An overlap-sensitive margin classifier for imbalanced and overlapping data. Expert Systems with Applications 98:72–83

82. Leyva E, González A, Perez R (2014) A set of complexity measures designed for applying meta-learning to instance selection. IEEE Transactions on Knowledge and Data Engineering 27(2):354–367
83. Li KS, Wang HR, Liu KH (2019) A novel error-correcting output codes algorithm based on genetic programming. Swarm and Evolutionary Computation 50:100564
84. Liu C (2008) Partial discriminative training for classification of overlapping classes in document analysis. International Journal of Document Analysis and Recognition (IJDAR) 11(2):53
85. Liu XY, Wu J, Zhou ZH (2008) Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39(2):539–550
86. López V, Fernández A, García S, Palade V, Herrera F (2013) An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences 250:113–141
87. Lorena AC, Costa IG, Spolaôr N, De Souto MC (2012) Analysis of complexity indices for classification problems: Cancer gene expression data. Neurocomputing 75(1):33–42
88. Lorena AC, Garcia LP, Lehmann J, Souto MC, Ho TK (2019) How complex is your classification problem? a survey on measuring classification complexity. ACM Computing Surveys (CSUR) 52(5):1–34
89. Luengo J, Fernández A, García S, Herrera F (2011) Addressing data complexity for imbalanced data sets: analysis of smote-based oversampling and evolutionary undersampling. Soft Computing 15(10):1909–1936
90. MacCuish J, MacCuish N (2010) Clustering in bioinformatics and drug discovery. CRC Press
91. Macià N, Bernadó-Mansilla E (2014) Towards uci+: a mindful repository design. Information Sciences 261:237–262
92. Malina W (2001) Two-parameter fisher criterion. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 31(4):629–636
93. Mani I, Zhang I (2003) knn approach to unbalanced data distributions: a case study involving information extraction. In: Proceedings of workshop on learning from imbalanced datasets, ICML United States, vol 126
94. Manukyan A, Ceyhan E (2016) Classification of imbalanced data with a geometric digraph family. The Journal of Machine Learning Research 17(1):6504–6543
95. Massie S, Craw S, Wiratunga N (2005) Complexity-guided case discovery for case based reasoning. In: AAAI, vol 5, pp 216–221
96. de Melo VV, Lorena AC (2018) Using complexity measures to evolve synthetic classification datasets. In: 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, pp 1–8
97. Menzies T, Butcher A, Cok D, Marcus A, Layman L, Shull F, Turhan B, Zimmermann T (2012) Local versus global lessons for defect prediction and effort estimation. IEEE Transactions on software engineering 39(6):822–834
98. Mercier M, Santos M, Abreu P, Soares C, Soares J, Santos J (2018) Analysing the footprint of classifiers in overlapped and imbalanced contexts. In: International Symposium on Intelligent Data Analysis, Springer, pp 200–212
99. Muñoz MA, Villanova L, Baatar D, Smith-Miles K (2018) Instance spaces for machine learning classification. Machine Learning 107(1):109–147

100. Napierala K, Stefanowski J (2016) Types of minority class examples and their influence on learning classifiers from imbalanced data. Journal of Intelligent Information Systems 46(3):563–597
101. Napierała K, Stefanowski J, Wilk S (2010) Learning from imbalanced data in presence of noisy and borderline examples. In: International Conference on Rough Sets and Current Trends in Computing, Springer, pp 158–167
102. Nekooeimehr I, Lai-Yuen SK (2016) Adaptive semi-unsupervised weighted oversampling (a-suwo) for imbalanced datasets. Expert Systems with Applications 46:405–416
103. Oh S (2011) A new dataset evaluation method based on category overlap. Computers in Biology and Medicine 41(2):115–122
104. Orriols-Puig A, Macia N, Ho TK (2010) Documentation for the data complexity library in c++. Universitat Ramon Llull, La Salle 196:1–40
105. Pascual-Triana JD, Charte D, Andrés Arroyo M, Fernández A, Herrera F (2021) Revisiting data complexity metrics based on morphology for overlap and imbalance: snapshot, new overlap number of balls metrics and singular problems prospect. Knowledge and Information Systems 63(7):1961–1989
106. Prati R, G B, Monard M (2004) Class imbalances versus class overlapping: an analysis of a learning system behavior. In: Mexican international conference on artificial intelligence, Springer, pp 312–321
107. Rivolli A, Garcia LP, Soares C, Vanschoren J, de Carvalho AC (2018) Characterizing classification datasets: A study of meta-features for meta-learning. arXiv preprint arXiv:180810406
108. Sáez J, Luengo J, Stefanowski J, Herrera F (2015) Smote–ipf: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. Information Sciences 291:184–203
109. Sáez JA, Galar M, Krawczyk B (2019) Addressing the overlapping data problem in classification using the one-vs-one decomposition strategy. IEEE Access 7:83396–83411
110. Santos M, Abreu P, García-Laencina P, Simão A, Carvalho A (2015) A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. Journal of biomedical informatics 58:49–59
111. Santos M, Soares J, Abreu P, Araújo H, Santos J (2018) Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches. IEEE Computational Intelligence Magazine 13(3):59–76
112. Santoso B, Wijayanto H, Notodiputro KA, Sartono B (2018) K-neighbor over-sampling with cleaning data: a new approach to improve classification performance in data sets with class imbalance. Applied Mathematical Sciences 12(10):449–460
113. Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A (2009) Rusboost: A hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 40(1):185–197
114. Selvaraj G, Kaliamurthi S, Kaushik A, Khan A, Wei Y, Cho W, Gu K, Wei D (2018) Identification of target gene and prognostic evaluation for lung adenocarcinoma using gene expression meta-analysis, network analysis and neural network algorithms. Journal of biomedical informatics 86:120–134
115. Shilaskar S, Ghatol A, Chatur P (2017) Medical decision support system for extremely imbalanced datasets. Information Sciences 384:205–219

116. Singh D, Gosain A, Saha A (2020) Weighted k-nearest neighbor based data complexity metrics for imbalanced datasets. Statistical Analysis and Data Mining: The ASA Data Science Journal 13(4):394–404
117. Singh S (2003) Multiresolution estimates of classification complexity. IEEE Transactions on pattern analysis and machine intelligence 25(12):1534–1539
118. Singh S (2003) Prism–a novel framework for pattern recognition. Pattern Analysis & Applications 6(2):134–149
119. Slowik A, Kwasnicka H (2020) Evolutionary algorithms and their applications to engineering problems. Neural Computing and Applications pp 1–17
120. Smith MR, Martinez T, Giraud-Carrier C (2014) An instance level analysis of data complexity. Machine learning 95(2):225–256
121. Sotoca JM, Sanchez J, Mollineda RA (2005) 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 pp 77–83
122. Sotoca JM, Mollineda RA, Sánchez JS (2006) A meta-learning framework for pattern classication by means of data complexity measures. Inteligencia Artificial Revista Iberoamericana de Inteligencia Artificial 10(29):31–38
123. Sowah RA, Agebure MA, Mills GA, Koumadi KM, Fiawoo SY (2016) New cluster undersampling technique for class imbalance learning. International Journal of Machine Learning and Computing 6(3):205
124. Stefanowski J (2013) Overlapping, rare examples and class decomposition in learning classifiers from imbalanced data. In: Emerging paradigms in machine learning, Springer, pp 277–306
125. Stefanowski J (2016) Dealing with data difficulty factors while learning from imbalanced data. In: Challenges in Computational Statistics and Data Mining, Springer, pp 333–363
126. Stefanowski J, Wilk S (2008) Selective pre-processing of imbalanced data for improving classification performance. In: International Conference on Data Warehousing and Knowledge Discovery, Springer, pp 283–292
127. Tang W, Mao K, Mak LO, Ng GW (2010) Classification for overlapping classes using optimized overlapping region detection and soft decision. In: 2010 13th International Conference on Information Fusion, IEEE, pp 1–8
128. Tang Y, Gao J (2007) Improved classification for problem involving overlapping patterns. IEICE Transactions on Information and Systems 90(11):1787–1795
129. Thornton C (1998) Separability is a learner's best friend. In: 4th Neural Computation and Psychology Workshop, London, 9–11 April 1997, Springer, pp 40–46
130. Tomek I (1976) Two modifications of cnn. IEEE Transactions on Systems Man and Communications 6:769–772
131. Vorraboot P, Rasmequan S, Chinnasarn K, Lursinsap C (2015) Improving classification rate constrained to imbalanced data between overlapped and non-overlapped regions by hybrid algorithms. Neurocomputing 152:429–443
132. Vuttipittayamongkol P, Elyan E (2020) Improved overlap-based undersampling for imbalanced dataset classification with application to epilepsy and parkinson's disease. International journal of neural systems 30(08):2050043
133. Vuttipittayamongkol P, Elyan E (2020) Neighbourhood-based undersampling approach for handling imbalanced and overlapped data. Information Sciences 509:47–70

134. Vuttipittayamongkol P, Elyan E, Petrovski A, Jayne C (2018) Overlap-based undersampling for improving imbalanced data classification. In: International Conference on Intelligent Data Engineering and Automated Learning, Springer, pp 689–697
135. Vuttipittayamongkol P, Elyan E, Petrovski A (2020) On the class overlap problem in imbalanced data classification. Knowledge-based systems p 106631
136. Van der Walt CM, Barnard E (2007) Measures for the characterisation of pattern-recognition data sets. 18th Annual Symposium of the Pattern Recognition Association of South Africa ...
137. Van der Walt CM, et al. (2008) Data measures that characterise classification problems. PhD thesis, University of Pretoria
138. Wang BX, Japkowicz N (2010) Boosting support vector machines for imbalanced data sets. Knowledge and information systems 25(1):1–20
139. Wang S, Yao X (2009) Diversity analysis on imbalanced data sets by using ensemble models. In: 2009 IEEE symposium on computational intelligence and data mining, IEEE, pp 324–331
140. Wang S, Yao X (2013) Using class imbalance learning for software defect prediction. IEEE Transactions on Reliability 62(2):434–443
141. Wei J, Huang H, Yao L, Hu Y, Fan Q, Huang D (2020) Ia-suwo: An improving adaptive semi-unsupervised weighted oversampling for imbalanced classification problems. Knowledge-Based Systems 203:106116
142. Wei J, Huang H, Yao L, Hu Y, Fan Q, Huang D (2020) Ni-mwmote: An improving noise-immunity majority weighted minority oversampling technique for imbalanced classification problems. Expert Systems with Applications 158:113504
143. Weng CG, Poon J (2006) A data complexity analysis on imbalanced datasets and an alternative imbalance recovering strategy. In: 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06), IEEE, pp 270–276
144. Wilson DL (1972) Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics (3):408–421
145. Wojciechowski S, Wilk S (2017) Difficulty factors and preprocessing in imbalanced data sets: an experimental study on artificial data. Foundations of Computing and Decision Sciences 42(2):149–176
146. Wozniak M, Grana M, Corchado E (2014) A survey of multiple classifier systems as hybrid systems. Information Fusion 16:3–17
147. Xiong H, Wu J, Liu L (2010) Classification with classoverlapping: A systematic study. In: Proceedings of the 1st International Conference on E-Business Intelligence (ICEBI2010),, Atlantis Press
148. Yan Y, Liu R, Ding Z, Du X, Chen J, Zhang Y (2019) A parameter-free cleaning method for smote in imbalanced classification. IEEE Access 7:23537–23548
149. Yen SJ, Lee YS (2009) Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications 36(3):5718–5727
150. Zhu C, Wang Z (2017) Entropy-based matrix learning machine for imbalanced data sets. Pattern Recognition Letters 88:72–80

151. Zhu T, Lin Y, Liu Y (2017) Synthetic minority oversampling technique for multiclass imbalance problems. Pattern Recognition 72:327–340
152. Zhu T, Lin Y, Liu Y (2020) Improving interpolation-based oversampling for imbalanced data learning. Knowledge-Based Systems 187:104826
153. Zhu Y, Yan Y, Zhang Y, Zhang Y (2020) Ehso: Evolutionary hybrid sampling in overlapping scenarios for imbalanced learning. Neurocomputing 417:333–346