Analysing the Footprint of Classifiers in Overlapped and Imbalanced Contexts

Marta Mercier[1], Miriam S. Santos[1], Pedro H. Abreu[1(✉)], Carlos Soares[2], Jastin P. Soares[1], and João Santos[3]

[1] CISUC, Department of Informatics Engineering, University of Coimbra, Coimbra, Portugal
{mmercier, miriams, jastinps}@student.dei.uc.pt, pha@dei.uc.pt
[2] INESC TEC, Faculty of Engineering, University of Porto, Porto, Portugal
csoares@fe.up.pt
[3] IPO-Porto Research Centre (CI-IPOP), Porto, Portugal
joao.santos@ipoporto.min-saude.pt

Abstract. It is recognised that the imbalanced data problem is aggravated by other difficulty factors, such as class overlap. Over the years, several research works have focused on this problematic, although presenting two major hitches: the limitation of test domains and the lack of a formulation of the overlap degree, which makes results hard to generalise. This work studies the performance degradation of classifiers with distinct learning biases in overlap and imbalanced contexts, focusing on the characteristics of the test domains (shape, dimensionality and imbalance ratio) and on to what extent our proposed overlapping measure (*degOver*) is aligned with the performance results observed. Our results show that MLP and CART classifiers are the most robust to high levels of class overlap, even for complex domains, and that KNN and linear SVM are the most aligned with *degOver*. Furthermore, we found that the dimensionality of data also plays an important role in explaining performance results.

Keywords: Imbalanced data · Class overlap · Machine learning classifiers

1 Introduction

Data imbalance occurs when there is a considerable difference between the class priors of a given problem and, for a binary classification scenario, is commonly described by the Imbalance Ratio, IR $= \frac{n_{maj}}{n_{min}}$, where n_{maj} and n_{min} represent the number of majority and minority examples in the domain [2]. Prediction models built from imbalanced datasets are most often biased towards the

This article is a result of the project NORTE-01-0145-FEDER-000027, supported by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF).

W. Duivesteijn et al. (Eds.): IDA 2018, LNCS 11191, pp. 200–212, 2018.
https://doi.org/10.1007/978-3-030-01768-2_17

majority concept [9], which is especially critical when there is a higher cost of misclassifying the minority examples, such as diagnosing rare diseases, preventing fraud or detecting faulty systems [4,14]. However, data imbalance is not the sole factor that affects the performance of classifiers. As stated in recent literature, there are several others that combined with data imbalance, create a rather chaotic setting [12]. These are frequently referred to as *data difficulty factors* and commonly include: class overlap, small data set size/lack of density, the presence of small disjuncts and the existence of different types of minority examples (e.g. safe, borderline, rare and outlier examples) [16].

The problem of class overlap in imbalanced domains has been previously discussed in related work, although not with the required depth. The main objective of related work is to show that class imbalance is not the sole factor that affects classification performance, and that overlap plays an important role as well. However, authors often fail to provide some insights on how both problems act together and affect well-established classifiers, and to what extent one problem is more critical than the other for different learning biases. Furthermore, related work is also limited in the following aspects:

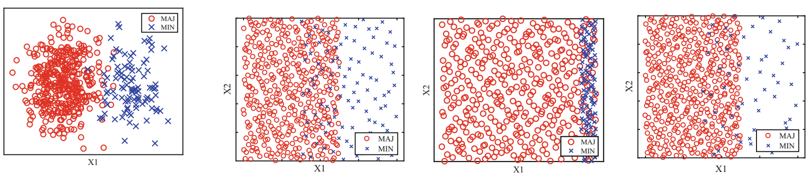
- **Definition of overlapping degree:** Some authors define the overlapping degree as a distance between minority and majority classes [3,10,13], which is only appropriate for specific data structures/shapes, while others define it as an intersection region of the majority and minority class, although without presenting a clear formulae to the define the degree of overlapping [5–7]. Other authors approximate the overlapping degree by considering the overlap of individual features (e.g. Fisher Discriminant Ratio – F1 measure) [11] or by identifying minority borderline examples [12,15], which may not completely capture the overall overlapping of the domains.
- **Tested domains:** Most research works consider artificial domains where the data structure is limited and unlikely to be found in real-world scenarios [5–7], besides being limited to two to five dimensions [3,13]. Others consider more complex shapes (e.g. linear versus non-linear shapes), however, limited to a two-dimensional space [12,15].
- **Nature of data and classifiers:** In the majority of works, only one or two/three classifiers are tested. The research of García et al. [5–7] is an exception, where different inductive biases are discussed, and it is possible to distinguish the behaviour of local versus global classifiers, although not in depth. Furthermore, performance results are most often discussed from a general perspective, rather than attending to the characteristics of the tested domains.

We have replicated several imbalanced scenarios with different characteristics found in related work and compare them altogether. These scenarios are generated for different degrees of imbalance and overlap, and the performance of standard classifiers is analysed. We also put an effort to fill in the gaps in related work by defining and evaluating a measure of the overlapping degree (*degOver*), considering artificial domains with different shapes and dimensionality (2–40 dimensions). Our experiments are focused on studying the behaviour of classifiers with distinct learning biases to determine whether some are more affected than others. This study is furthermore taken from different perspectives: focusing on the properties of the tested domains (shape and dimensionality), and focusing on to what extent the proposed overlapping measure is aligned with the performance results of the studied classifiers.

2 Related Work

The work of Prati et al. was one of the first studies on the impact of overlap in imbalanced domains [13]. Their domains consisted of two 5-dimensional clusters (Fig. 1a), where the distribution of minority and majority examples, as well as the distance between cluster centroids, could be changed (1–9 standard deviations). The classification results (C4.5) showed that the influence of the degree of imbalance becomes weaker as the distance between centroids increases.

García et al. [7] performed a similar experiment with 2-dimensional domains, where the majority and minority classes start well-separated and, for a fixed IR, the majority class moves towards the minority class, increasing the amount of overlap (Fig. 1b). Similarly to Prati et al. [13], authors concluded that the increasing overlap deteriorated the performance of classifiers. In a later work [5], authors distinguish between typical and atypical domains (Figs. 1b/d and c, respectively).



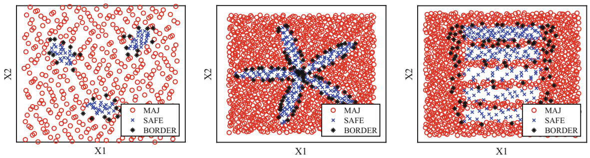
(a) (b) (c) (d)

Fig. 1. Artificial domains generated by Prati et al. [13] (a) and García et al. [7] (b–d).

Authors found that for typical domains, classifiers with a local nature (e.g. KNN) were more subjected to loss in performance for the majority class than classifiers with a more global learning. Regarding atypical domains, the classification results suggested that the recognition rate of the minority class improved as the minority class became denser. Denil and Trappenberg [3] also studied the joint-effect of class imbalance and overlap: they generated two-dimensional domains where both the class overlap and class imbalance could be changed. Their analysis was focused on the performance of SVM, showing that as the training size increases, the influence of class imbalance is negligible and that overlap is the main responsible for performance degradation.

The research of Luengo et al. [11] was not focused on the effects of class imbalance and overlap, although authors found that one measure of overlap between classes (F1 measure) proved to be informative of good/bad behaviour of classifiers.

Finally, we refer to the line of research of Napierala and Stefanowski [12,15], where class overlap is defined via the percentage of borderline minority examples. Napierala and Stefanowski studied the influence of disturbing minority class borders in three different 2-dimensional domains with different characteristics, *paw*, *clover* and *subclus*, (Fig. 2) and concluded that increasingly adding borderline examples degraded the classification performance [12].



(a) (b) (c)

Fig. 2. Artificial domains generated according to Napierala and Stefanowski [12,15].

As stated in the Introduction, a common limitation of related work is in the way class overlap is measured. In the research work of Prati et al. [13], increasing the distance between cluster centroids guarantees that the overlap is being reduced, although it is not possible to quantify the exact degree of overlap in each configuration. In García et al. [5–7], authors generate an artificial domain represented by a square of length 100 where both classes are defined uniformly in a rectangle of 50×100 (typical domain). The IR was fixed to 4:1, while the overlapping degree was controlled through the distance between the square centres. Initially, the majority and minority squares start well separated by a line orthogonal to $X1$ axis, and increasing amounts of class overlap are produced by moving the majority square towards the minority square in a stepwise manner: [0..50], [10..60], [20..70], [30..80], [40..90] and [50..100] for 0, 20, 40, 60, 80 and 100% overlap. Let us consider the example given in Fig. 1b, for a typical domain with IR 4:1 and 40% overlap. Since no formulae is presented in the original papers [5–7], we may assume that the calculation of the overlap degree was performed as a fraction of the area that is overlapped (A_{inter}) over the total minority area (A_{min}) (or majority area, since they are equal). In that way we would obtain $overlap = \frac{A_{inter}}{A_{min}} = \frac{2000}{5000} = 40\%$. If the IR was defined arbitrarily, then Fig. 1d, with an IR of 8:1, would also illustrate a scenario with 40% class overlap. However, if we consider the definition of class overlap as regions in the data space with similar priors [10], this does not seem correct, since the number of points that occupies the same region is lower in Fig. 1d. Basing our reasoning on the similarity of class priors, a 8:1 configuration should produce a lower degree of overlap.

For atypical situations, the majority examples are always uniformly distributed in a square of length 100, while the minority examples are condensed in ranges [75..100], [80..100], [85..100], [90..100] and [95..100]. If the same rationale as above is applied to atypical domains (Fig. 1c), the percentage of overlap would be 100%, since the minority area is completely embedded in the majority area. In their paper [5], authors do not elaborate on the percentage of overlap present in each configuration – no percentages or any other values are presented for the overlapping amount. Instead, these domains are evaluated in terms of global imbalance, local imbalance and the size of the overlapping region: the notion of overlap gets somewhat lost, which complicated the discussion of results. According to the definition of class overlap as "regions in the data space with similar priors" [3,10], we believe that the "local imbalance in the overlap region" implies the existence of an overlap degree. For instance, since the distribution of examples is uniform, a [75..100] range of minority examples over the majority class square means that both classes have the same number of patterns (100 points of each class) in the overlap region, thus, there is no local imbalance in the overlap region. In this situation, the priors of both classes are the same, and therefore the overlap degree should be maximum. As the minority class becomes denser, the local imbalance increases because the size of the overlapping region is decreased, meaning that the class priors are uneven, and therefore the overlap, in fact, is decreasing. From this perspective, we could evaluate the results as follows: as the minority class becomes denser, the overlapping degree is decreasing and therefore the classification performance improves.

Regarding the F1 measure used in the research of Luengo et al. [11], it measures the highest discriminative power in all the features in the data. Essentially, F1 is measured for all the features in the dataset according to $F1 = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$, and the highest value among all features is returned. Therefore, F1 measures the overlapping of individual features, not the "overall overlapping of data". If two domains have the same structure (features have the same range and spread), F1 assumes the same or similar value, although they might be different in classification terms.

Finally, regarding the typology defined by Napierala and Stefanowski [12], as only the minority class is considered, borderline examples from the majority class (that contribute to class overlapping) are not identified. Also, as the percentage is determined over the total minority examples, majority regions where there are no examples from the minority class are not taken into account.

An overlapping degree should attend to regions with the same class priors (rather than considering distances between classes or the size of overlapping areas only), consider the overall overlap (rather than the overlap of individual features or focusing solely on the minority class examples) and focus on the characteristics of data space: structure and class decomposition, distribution of examples (implying that class imbalance could affect class overlap) and data dimensionality. In a recent work, Lee and Kim propose a hybrid classifier based on a fuzzy support vector machine and k-nearest neighbour algorithm to address class imbalance and overlapping simultaneously [8]: the data space is divided into soft and hard overlap regions so that each is handled separately. Although the focus of the work is not to analyse the joint-impact of these problems, authors define overlap-sensitive costs, where each example is classified as being part of an overlapping or a non-overlapping region, through a k neighbourhood-based function. This approach is advantageous since it considers the factors mentioned above and therefore we have decided to adapt it in order to formulate a degree of overlap and analyse its behaviour when applied to several data characteristics and imbalance ratios.

3 Experiments

All datasets contained 1500 examples and were generated with increasing levels of imbalance, namely 1:1, 2:1, 4:1, 6:1, 8:1 and 10:1, and increasing number of dimensions, namely 2, 3, 5, 10, 15, 20, 30 and 40D. The datasets further considered several overlap degrees and data structures (shapes), resulting in different levels of complexity for classifiers: *clusters* and *garcia* (less complex shapes) and *clover*, *paw* and *subclus* (more complex shapes).

The *clusters* domains, *clusters-vo* (Fig. 1a) and *clusters-va*, consist of two normal distributions (one for each class) where each cluster has unitary standard deviation. For *clusters-vo* only one of the attributes is changed and the overlap region decreases as the separation in the $X1$ axis between cluster centres increases. For *clusters-va*, all the attributes are changed and the separation is increased in all axis, according to the number of dimensions. The *garcia* domains, *garcia-va* and *garcia-vo* (Fig. 1b), follow a rectangular shape where both class are centred in the same point, being overlapped. The distance between the centres is then increased in steps of 10 units until 3×radius for *garcia-va* or 4×radius for *garcia-vo* is reached, guaranteeing no overlap. The *paw*, *clover* and *subclus* scenarios (Figs. 2a, b and c, respectively) are composed by different shapes of the minority class, and the remaining space is filled by the majority class. The minority class is formed by two types of examples – safe (located in homogeneous regions of the class) and borderline (located in the boundary between both classes). For each imbalance ratio and dimension, the ratio of safe/borderline examples varies from 100/0 to 0/100.

We measured the degree of overlap using a neighbourhood function. For each example x_i in data (considering both classes), its 5-nearest neighbours are found: if x_i and all its 5-nearest neighbours are from the same class, then example x_i belong to a non-overlapping region; otherwise, it belongs to an overlapping region. The number of examples (considering both classes) that belong to overlapping regions (n_{min_over} and n_{maj_over}) are then divided by the total number of examples, n. Thus, $degOver = (n_{min\_over} + n_{maj\_over})/n$ measures the percentage of examples comprised in overlapping regions. Measuring the degree of overlap as a neighbourhood-based function has two main advantages: it can be applied to d-dimensional data with different structures/shapes and takes the imbalance ratio (IR) into account. Besides considering the IR as a fraction of n_{maj}/n_{min}, we have normalised this ratio to measure the severity of the imbalance ratio. The degree of imbalance is defined as $degIR = 1 - \frac{n_{min}}{n/2}$. The value of n_{min} is naturally affected by the IR, and for a particular IR (e.g. $IR = 4$) and total number of examples (e.g. $n = 500$), is computed as $n_{min} = n/(IR + 1)$, (for IR = 4:1, $n_{min} = 500/(4+1) = 100$ minority examples and $degIR = 1 - 100/(500/2) = 0.6$). This degree of imbalance reflects how much a particular scenario is affected by class imbalance on a normalised scale between 0 and 1. We analysed seven classifiers with distinct inductive biases [1]: Classification and Regression Trees (CART), k-Nearest Neighbour (KNN), Fisher Linear Discriminant (FLD), Naive Bayes Classifier (NB), Multilayer Perceptron (MLP), Support Vector Machine with a linear kernel (SVM-linear) and Support Vector Machine with radial basis kernel (SVM-rbf). Regarding the evaluation of the classification performance, similarly to previous work [5,16], we use Sensitivity (SENS) and Specificity (SPEC).

4 Results and Discussion

We start by analysing the performance degradation of each classifier according to the properties of the test domains (IR, structure/shape and dimensionality). To analyse this degradation, we first tuned the parameters of all classifiers (k for KNN, C for SVM-linear, C and γ for SVM-rbf and number of neurons and layers for MLP) on the configuration with the least amount of overlap, for each domain, IR and dimensionality. Then, we analysed how much the defined model is affected by increasing levels of overlap. The Sensitivity results for the minority class are presented in Table 1, as well as the *degOver* for all the presented domains (due to space restrictions, we report only the Sensitivity, although the Specificity was analysed as well). Overall, CART, MLP and KNN show the lowest degradation in classification performance (considering both Sensitivity and Specificity) for all the test domains, whereas FLD and SVM-linear suffer the most with the increase of class overlap. These latter two classifiers also seem to be critically affected by the IR and data structure: the Sensitivity of FLD becomes 0 for 4:1 ratios and higher (*clover* and *subclus* domains), while SVM-linear struggles with both higher IR and higher dimensions (for *clover* and *subclus*) with Sensitivity results of 0 for ratios higher than 4:1 in higher dimensions (15 and 40D). Thus, linear classifiers seem to be affected by all four components of the problem (IR, dimensionality, class overlap and data structure), where the data structure seems to be the most prominent factor.

CART, MLP and KNN, although with different classification paradigms, are able to "adapt" to the data structure more easily, handling data that is not linearly separable: CART by recursively partitioning the input space, MLP by using multiple layers with non-linear activation functions and KNN through its neighbourhood function. These three classifiers have only achieved a poor performance for *clusters-va* and *garcia-va*, when both clusters/squares are centred at the same coordinates, respectively. These poor results are consistent with higher values of *degOver* (between 0.4 and 0.97), although *degOver* is not capable of explaining this effect entirely: in *clover* and *subclus* domains, there are some scenarios achieving the same overlapping values, where KNN, MLP and CART perform well. This may be mostly due to the structure of the domain

Table 1. Sensitivity of classifiers for different domains, overlap levels, IR and dimensionality.

		degOver				CART				FLD				SVM-linear				SVM-rbf				NB				MLP				KNN			
Dimensions	**Overlap**	**1:1**	**4:1**	**6:1**	**10:1**	**1:1**	**4:1**	**6:1**	**10:1**	**1:1**	**4:1**	**6:1**	**10:1**	**1:1**	**4:1**	**6:1**	**10:1**	**1:1**	**4:1**	**6:1**	**10:1**	**1:1**	**4:1**	**6:1**	**10:1**	**1:1**	**4:1**	**6:1**	**10:1**	**1:1**	**4:1**	**6:1**	**10:1**
clover																																	
2D	S100-B0	0.201	0.221	0.205	0.162	0.91	0.78	0.70	0.66	0.48	0.00	0.00	0.00	0.69	0.47	0.50	0.46	1.00	0.95	0.93	0.88	0.91	0.23	0.05	0.00	0.99	0.93	0.91	0.79	1.00	0.91	0.81	0.66
	S50-B50	0.261	0.301	0.274	0.214	0.94	0.73	0.67	0.36	0.47	0.00	0.00	0.00	0.66	0.31	0.41	0.40	1.00	0.92	0.85	0.66	0.88	0.18	0.05	0.00	0.99	0.91	0.86	0.74	1.00	0.79	0.65	0.49
	S0-B100	0.290	0.339	0.277	0.228	0.94	0.71	0.53	0.44	0.48	0.00	0.00	0.00	0.68	0.31	0.46	0.44	1.00	0.97	0.93	0.73	0.88	0.26	0.10	0.00	1.00	0.94	0.90	0.79	1.00	0.80	0.62	0.39
5D	S100-B0	0.044	0.047	0.040	0.042	0.99	0.98	0.99	0.94	0.51	0.00	0.00	0.00	0.97	0.36	0.43	0.15	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.91	0.99	1.00	1.00	1.00	1.00
	S50-B50	0.043	0.043	0.046	0.044	1.00	0.98	0.96	0.96	0.52	0.00	0.00	0.00	0.93	0.21	0.20	0.19	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.99	1.00	1.00	1.00	0.99
	S0-B100	0.045	0.046	0.045	0.052	1.00	0.97	0.97	0.93	0.48	0.00	0.00	0.00	0.98	0.37	0.19	0.39	1.00	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.98	1.00	1.00	1.00	0.99
15D	S100-B0	0.140	0.095	0.089	0.082	1.00	0.99	0.96	0.95	0.52	0.00	0.00	0.00	1.00	0.06	0.00	0.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	S50-B50	0.148	0.090	0.081	0.073	1.00	0.98	0.99	0.96	0.53	0.00	0.00	0.00	1.00	0.05	0.00	0.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	S0-B100	0.140	0.098	0.083	0.079	0.99	0.96	0.96	0.93	0.50	0.00	0.00	0.00	1.00	0.05	0.00	0.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
40d	S100-B0	0.499	0.095	0.089	0.082	1.00	1.00	1.00	1.00	0.92	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	S50-B50	0.499	0.090	0.081	0.073	1.00	0.99	1.00	0.99	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	S0-B100	0.500	0.098	0.083	0.079	1.00	0.99	0.99	0.99	0.65	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
subclus																																	
2D	S100-B0	0.202	0.230	0.208	0.169	1.00	0.98	0.97	0.93	0.63	0.00	0.00	0.00	0.74	0.48	0.38	0.33	1.00	0.90	0.78	0.69	1.00	0.53	0.27	0.00	0.98	0.80	0.88	0.81	1.00	0.85	0.79	0.65
	S50-B50	0.340	0.317	0.266	0.229	0.98	0.90	0.86	0.73	0.65	0.00	0.00	0.00	0.72	0.40	0.31	0.12	0.00	0.85	0.69	0.54	1.00	0.46	0.10	0.00	0.97	0.00	0.67	0.57	0.98	0.66	0.56	0.48
	S0-B100	0.353	0.340	0.307	0.248	0.97	0.89	0.85	0.70	0.67	0.00	0.00	0.00	0.69	0.24	0.22	0.00	0.01	0.79	0.71	0.36	1.00	0.35	0.02	0.00	0.90	0.00	0.63	0.49	0.99	0.66	0.45	0.35
5D	S100-B0	0.251	0.253	0.236	0.199	1.00	0.97	0.96	0.96	0.64	0.00	0.00	0.00	0.73	0.48	0.48	0.41	1.00	0.96	0.92	0.84	1.00	1.00	1.00	0.96	0.99	0.89	0.88	0.77	1.00	0.99	0.96	0.83
	S50-B50	0.300	0.283	0.259	0.219	1.00	0.97	0.97	0.89	0.64	0.00	0.00	0.00	0.72	0.47	0.42	0.35	1.00	0.94	0.90	0.75	1.00	1.00	0.99	0.93	0.98	0.83	0.86	0.69	0.99	0.97	0.93	0.78
	S0-B100	0.295	0.315	0.303	0.241	0.99	0.98	0.96	0.90	0.65	0.00	0.00	0.00	0.70	0.40	0.49	0.34	1.00	0.94	0.89	0.79	1.00	1.00	1.00	0.82	0.98	0.81	0.75	0.65	1.00	0.94	0.83	0.69
15D	S100-B0	0.493	0.661	0.625	0.575	1.00	0.98	0.95	0.93	0.64	0.00	0.00	0.00	0.72	0.08	0.07	0.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00	0.99	0.96	0.94	0.93	1.00	1.00	1.00	1.00
	S50-B50	0.495	0.671	0.655	0.593	1.00	0.99	0.96	0.90	0.64	0.00	0.00	0.00	0.72	0.09	0.08	0.00	1.00	0.99	0.97	0.99	1.00	1.00	1.00	1.00	0.98	0.92	0.96	0.84	1.00	1.00	1.00	0.99
	S0-B100	0.494	0.677	0.661	0.599	1.00	0.99	0.99	0.95	0.63	0.00	0.00	0.00	0.73	0.12	0.07	0.00	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.93	0.91	0.79	1.00	1.00	1.00	0.99
40D	S100-B0	0.500	0.800	0.857	0.909	1.00	0.98	0.96	0.90	0.64	0.00	0.00	0.00	0.72	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.96	1.00	1.00	1.00	1.00
	S50-B50	0.500	0.800	0.857	0.909	1.00	0.95	0.92	0.93	0.64	0.00	0.00	0.00	0.71	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.98	0.98	1.00	1.00	1.00	1.00
	S0-B100	0.500	0.800	0.857	0.909	1.00	0.99	0.98	0.92	0.64	0.00	0.00	0.00	0.70	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.96	1.00	1.00	1.00	1.00
clusters-va																																	
2D	8 std	0.000	0.000	0.000	0.000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.43	0.72	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.90	1.00	1.00	1.00	1.00
	4 std	0.007	0.007	0.005	0.001	1.00	0.99	0.99	0.98	1.00	0.99	1.00	1.00	1.00	0.99	0.99	0.96	0.23	0.44	0.00	0.00	1.00	0.99	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.99	1.00	1.00
	0 std	0.961	0.741	0.580	0.427	0.51	0.18	0.08	0.08	0.49	0.00	0.00	0.00	0.49	0.00	0.00	0.00	0.86	0.00	0.00	0.00	0.45	0.00	0.00	0.00	0.15	0.01	0.01	0.00	0.50	0.20	0.15	0.09
5D	8 std	0.000	0.000	0.000	0.000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	4 std	0.000	0.000	0.000	0.000	1.00	0.99	0.98	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	0.25	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0 std	0.971	0.731	0.598	0.448	0.49	0.17	0.13	0.05	0.52	0.00	0.00	0.00	0.48	0.00	0.00	0.00	0.22	0.00	0.00	0.00	0.59	0.00	0.01	0.00	0.46	0.02	0.04	0.01	0.50	0.17	0.14	0.10
15D	8 std	0.000	0.000	0.000	0.000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.70	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	4 std	0.000	0.000	0.000	0.000	1.00	0.98	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0 std	0.973	0.735	0.603	0.416	0.47	0.22	0.21	0.11	0.50	0.00	0.00	0.00	0.39	0.00	0.00	0.00	0.80	0.00	0.00	0.00	0.48	0.00	0.01	0.02	0.45	0.15	0.13	0.09	0.50	0.22	0.13	0.07
40D	8 std	0.000	0.000	0.000	0.000	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	4 std	0.000	0.000	0.000	0.000	0.99	0.99	0.96	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0 std	0.961	0.751	0.603	0.445	0.52	0.22	0.16	0.08	0.50	0.00	0.00	0.00	0.54	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.52	0.05	0.02	0.01	0.48	0.24	0.14	0.11	0.49	0.23	0.12	0.10
garcia-va																																	
2d	step-100	0.000	0.000	0.000	0.000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	step-50	0.106	0.089	0.081	0.056	0.94	0.91	0.91	0.91	0.96	0.90	0.87	0.84	0.96	0.90	0.87	0.83	0.94	0.00	0.00	0.84	0.90	0.87	0.86	0.83	0.96	0.90	0.89	0.86	0.93	0.90	0.90	0.88
	step-0	1.000	0.887	0.775	0.555	0.20	0.01	0.00	0.00	0.46	0.00	0.00	0.00	0.46	0.00	0.00	0.00	0.29	0.00	0.00	0.00	0.43	0.00	0.00	0.00	0.28	0.00	0.00	0.00	0.23	0.04	0.00	0.00
5d	step-100	0.000	0.000	0.000	0.000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	step-50	0.011	0.013	0.007	0.005	1.00	0.98	0.99	0.96	1.00	1.00	0.99	0.99	1.00	1.00	0.99	0.99	1.00	0.93	0.47	0.00	1.00	1.00	0.99	0.99	1.00	0.99	0.99	0.98	1.00	0.99	1.00	0.98
	step-0	0.999	0.826	0.691	0.547	0.33	0.05	0.02	0.01	0.44	0.00	0.00	0.00	0.43	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.35	0.00	0.00	0.00	0.45	0.00	0.01	0.00	0.19	0.05	0.03	0.01
15d	step-100	0.000	0.000	0.000	0.000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	step-50	0.000	0.000	0.000	0.000	1.00	0.98	0.95	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.14	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	step-0	0.993	0.800	0.631	0.500	0.51	0.14	0.03	0.05	0.36	0.00	0.00	0.00	0.42	0.00	0.00	0.00	0.36	0.00	0.00	0.00	0.24	0.00	0.00	0.00	0.91	0.55	0.38	0.38	0.54	0.12	0.09	0.08
40d	step-100	0.000	0.000	0.000	0.000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	step-50	0.000	0.000	0.000	0.000	1.00	1.00	0.98	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	step-0	0.911	0.775	0.554	0.399	0.80	0.50	0.52	0.28	0.34	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.57	0.00	0.00	0.00	0.18	0.00	0.00	0.00	0.98	0.62	0.74	0.53	0.58	0.20	0.24	0.25

and the way overlap is generated. In *clover* and *subclus*, the structure of data is not changed (when the overlap increases, more borderline examples are added, but the core structure of the domain remains the same – Fig. 2). In *clusters-va* and *garcia-va*, the structure of data changes with the increase of overlap, since the cluster/square centres become closer (Figs. 1a and b). For these particular scenarios (*clusters-va* and *garcia-va*) it is also noticeable that higher IR ratios deteriorate the classification performance (for all dimensions), which is not observed for the remaining domains. Therefore, it is not possible to infer clearly what is more severe for these classifiers, since they seem to be affected by a combination of data structure, IR and class overlap, though not as severely by data dimensionality. Finally, although SVM-rbf seems to be more affected by class overlap than class imbalance, where the decrease in Sensitivity results was especially noticed in lower dimensions (2D) for most scenarios. Kernel methods are known to ease non-linear problems by mapping the input data to an "improved" feature space, but this largely depends on data itself. Furthermore, we also observed that this classifier has obtained poor Specificity results: it was not possible to define a clear decision hyperplane without compromising the classification of the majority class. On the contrary, NB suffered the most from higher IR, which is consistent with its bias to favour the most prevalent class, adjusting its decision threshold accordingly.

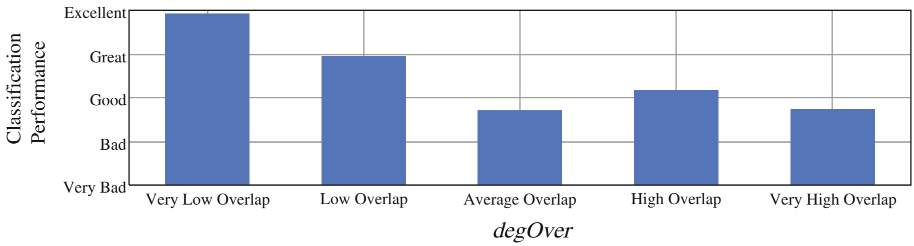


Fig. 3. Alignment between *degOver* and classification performance of KNN.

We now perform an analysis on the alignment of *degOver* and classification performance. As previously discussed, *degOver* may not be able to fully characterise the behaviour of all classifiers, although it may provide interesting insights in some cases. Of note is the ability of *degOver* to "adapt" to different IR levels: class overlap is not measured independently of class imbalance, and *degOver* generally assumes lower values as the IR increases, as discussed in Sect. 2. An exception occurs for the *subclus* domain for higher dimensions (15 and 40D), which shows that both the shape of domain and dimensionality may impact the results in certain scenarios. We then transformed *degOver* and classification performance to categories to ease the interpretation of results: *degOver* values were divided in five intervals from 0 to 1: *very low overlap* (VLO), *low overlap* (LO), *average overlap* (AO), *high overlap* (HO) and *very high overlap* (VHO),

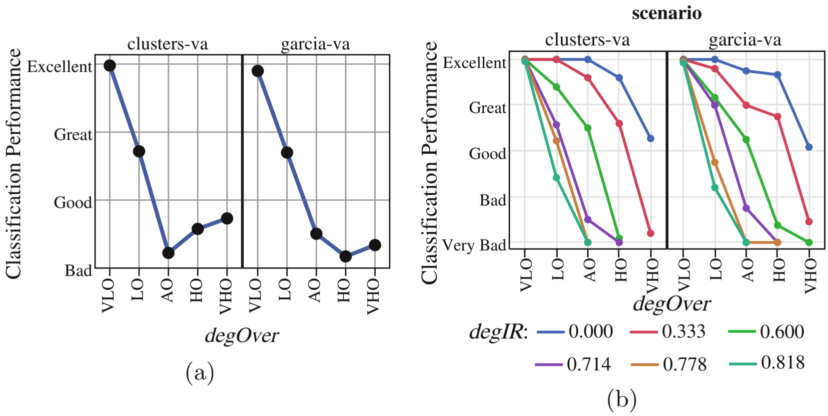
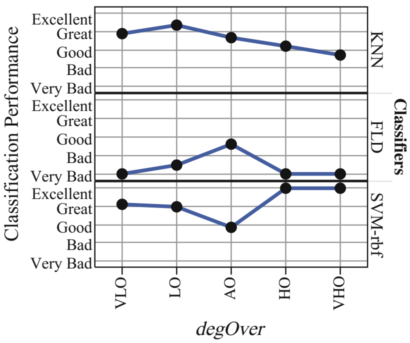


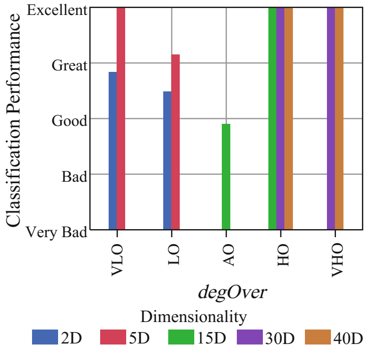
Fig. 4. (a) Alignment between *degOver* and classification performance of FLD (*clusters* and *garcia*); (b) Lines representing different levels of *degIR*.

while Sensitivity and Specificity results were combined to produce also five categories of performance: *very bad*, *bad*, *good*, *great* and *excellent*. Figure 3 shows the relationship between *degOver* and KNN, which was found to be the classifier most aligned with *degOver* (as expected, since their underlying principles are the same – they are based on neighbourhood functions). Overall, the performance of classifiers deteriorates with higher values of *degOver*, although this decrease is not linear: maximum levels of overlap do not necessarily correspond to minimum performance results. This suggests, as previously discussed, that there are other factors (namely, data structure) affecting the performance of classifiers, as will be discussed in what follows.

For *clusters* and *garcia*, classification performance and *degOver* are aligned for all classifiers: an example of this alignment is presented in Fig. 4a for FLD. The slight increase in performance for higher *degOver* values (HO and VHO) may be explained by the IR values (Fig. 4b): the blue line (normalised IR of 0) indicates that there is no class imbalance – in this scenario, although the overlap is high, the performance results are also high, causing the slight increase of performance for the high overlap levels in Fig. 4a. Again, these results suggest that all these properties of data (IR, class overlap and data structure) should be analysed together to better understand the performance of classifiers. For more complex scenarios, as *clover*, *subclus* and *paw*, the alignment with *degOver* varies for different classifiers. None of them presents the expected behaviour (a performance decrease for higher values of *degOver*) for all three domains, although KNN and SVM-linear present a better alignment than the remaining classifiers, being KNN clearly the most aligned (Fig. 5a). Figure 5a also presents the results for FLD and SVM-rbf, two of the classifiers that do not present a good alignment between *degOver* and classification performance for complex domains. We hypothesise that this mismatch can be related to the structure of data, which may be influenced by data dimensionality. Some classifiers (SVM-



(a)



(b)

Fig. 5. (a) Alignment between *degOver* and classification performance of KNN, FLD and SVM-rbf considering only more complex shapes (*paw*, *clover* and *subclus*); (b) Dimensionality discrimination for SVM-rbf.

rbf, CART, NB and MLP) are able to classify datasets with higher overlap levels in higher dimensions (Fig. 5b): the *subclus* domain is such an example (Table 1) where high *degOver* values occur in 40D, and the mentioned classifiers obtain better results than for lower dimensions (sometimes with lower *degOver* values as well). These results suggest that data dimensionality is especially relevant for more complex domains and that *degOver* may have to be adjusted according to the number of dimensions and number of examples in data in order to give more insights on properties of the domain.

5 Conclusions

Class overlap is one of the difficulty factors that deteriorates the performance of classifiers and is even more critical in imbalanced contexts, as discussed in related work. However, most authors study class overlap without providing a clear formula to measure its degree: overlap is often perceived as a distance between majority and minority concepts or as an area of intersection between majority and minority classes, without considering the IR nor the structure of data, which may limit the conclusions derived from such setups. From our perspective, a measure of the degree of overlap should take the IR and structure of data into account. Therefore, we evaluate the usefulness of *degOver* to quantify the overlapping degree and its relationship with the classification performance of standard classifiers in several domains with different shapes, IRs and dimensionality. Our results revealed that MLP and CART are less prone to suffer from high levels of overlap and show good performance even in the presence of more complex domains. Furthermore, in simpler scenarios, *degOver* is aligned with classification performance for all classifiers, even for varying amounts of imbalance. However, this alignment varies significantly in more complex domains and seems to be influenced by data dimensionality. In sum, although *degOver* takes the imbalance ratio into account and can be measured for any data structure and dimensionality, it needs to be adjusted to better represent these properties of data so that it may provide more useful insights for more complex domains.

References

1. Abreu, P.H., Santos, M.S., Abreu, M.H., Andrade, B., Silva, D.C.: Predicting breast cancer recurrence using machine learning techniques: a systematic review. ACM Comput. Surv. (CSUR) **49**(3), 52 (2016)
2. Ali, A., Shamsuddin, S.M., Ralescu, A.L.: Classification with class imbalance problem: a review. Int. J. Adv. Soft Compu. Appl. **7**(3), 176–204 (2015)
3. Denil, M., Trappenberg, T.: Overlap versus imbalance. In: Farzindar, A., Kešelj, V. (eds.) AI 2010. LNCS (LNAI), vol. 6085, pp. 220–231. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13059-5_22
4. Domingues, I., Amorim, J.P., Abreu, P.H., Duarte, H., Santos, J.: Evaluation of oversampling data balancing techniques in the context of ordinal classification, pp. 5691–5698 (2018)
5. García, V., Mollineda, R.A., Sánchez, J.S.: On the k-nn performance in a challenging scenario of imbalance and overlapping. Pattern Anal. Appl. **11**(3–4), 269–280 (2008)
6. García, V., Mollineda, R.A., Sánchez, J.S., Alejo, R., Sotoca, J.M.: When overlapping unexpectedly alters the class imbalance effects. In: Martí, J., Benedí, J.M., Mendonça, A.M., Serrat, J. (eds.) IbPRIA 2007. LNCS, vol. 4478, pp. 499–506. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-72849-8_63
7. García, V., Sánchez, J., Mollineda, R.: An empirical study of the behavior of classifiers on imbalanced and overlapped data sets. In: Rueda, L., Mery, D., Kittler, J. (eds.) CIARP 2007. LNCS, vol. 4756, pp. 397–406. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76725-1_42
8. Lee, H.K., Kim, S.B.: An overlap-sensitive margin classifier for imbalanced and overlapping data. Expert Syst. Appl. **98**, 72–83 (2018)
9. Longadge, R., Dongre, S.: Class imbalance problem in data mining review. arXiv preprint arXiv:1305.1707 (2013)
10. López, V., Fernández, A., García, S., Palade, V., Herrera, F.: An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inf. Sci. **250**, 113–141 (2013)
11. Luengo, J., Fernández, A., García, S., Herrera, F.: Addressing data complexity for imbalanced data sets: analysis of smote-based oversampling and evolutionary undersampling. Soft Comput. **15**(10), 1909–1936 (2011)
12. Napierała, K., Stefanowski, J., Wilk, S.: Learning from imbalanced data in presence of noisy and borderline examples. In: Szczuka, M., Kryszkiewicz, M., Ramanna, S., Jensen, R., Hu, Q. (eds.) RSCTC 2010. LNCS (LNAI), vol. 6086, pp. 158–167. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13529-3_18
13. Prati, R.C., Batista, G.E.A.P.A., Monard, M.C.: Class imbalances *versus* class overlapping: an analysis of a learning system behavior. In: Monroy, R., Arroyo-Figueroa, G., Sucar, L.E., Sossa, H. (eds.) MICAI 2004. LNCS (LNAI), vol. 2972, pp. 312–321. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24694-7_32

14. Santos, M.S., Abreu, P.H., García-Laencina, P.J., Simão, A., Carvalho, A.: A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. J. Biomed. Inf. **58**, 49–59 (2015)
15. Stefanowski, J.: Overlapping, rare examples and class decomposition in learning classifiers from imbalanced data. Emerging Paradigms in Machine Learning, pp. 277–306. Springer, Berlin (2013). https://doi.org/10.1007/978-3-642-28699-5_11
16. Wojciechowski, S., Wilk, S.: Difficulty factors and preprocessing in imbalanced data sets: an experimental study on artificial data. Found. Comput. Decis. Sci. **42**(2), 149–176 (2017)