Investigating the Performance of Data Complexity & Instance Hardness Measures as A Meta-Feature in Overlapping Classes Problem

Omaimah Saif Al Hosni*
School of Engineering, University of Aberdeen,
Scotland/UK, Email:
o.alhosni.19@abdn.ac.uk

Andrew Starkey
School of Engineering, University of Aberdeen,
Scotland/UK, Email:
a.starkey@abdn.ac.uk

ABSTRACT

Since the meta-learning recommendation's quality depends on the meta-features decision quality, a common problem in meta-learning is establishing a (good) collection of meta-features that best represent the dataset properties. Therefore, many meta-feature measures/methods have been proposed during the last decade to describe the characteristics of the data. However, little attention has been paid to validating the meta-feature decisions in reflecting the actual data properties. In particular, if the meta-feature analysis is negatively affected by complex data characteristics, such as class overlap due to the distortion imposed by the noisy features at the decision boundary of the classes and thereby produces biased meta-learning recommendations that do not match the actual data characteristics (either by overestimating or underestimating the complexity). Hence, this issue is crucial to ensure the success of the meta-learning model since the learning algorithm selection decision is based on meta-feature analysis. Based on that, in this work, we aim to investigate this by assessing the performance of Complexity Measures (global/data-level measures) & Instance Hardness Measures (local/instance-level measures) as a meta-feature in reflecting the actual data complexity associated with the high-class overlapping problem. The reason for focusing on the overlapping classes problem is that several studies have proven that this data issue significantly contributes to degrading prediction accuracy, with which most real-world datasets are associated. On the other hand, the motivation for using the above measures among different meta-feature methods proposed in the literature is that since this study aims to focus on the overlapping classes problem, the above measures are mainly proposed to estimate the data complexity according to the geometrical descriptions focusing on the class overlap imposed by feature values, in which match the data problem that the study interested to investigate.

CCS CONCEPTS

• **General and reference Cross-computing tools and techniques**; • **Computing methodologies** → Machine learning; Machine learning approaches; Instance-based learning;

This work is licensed under a Creative Commons Attribution International 4.0 License.

ICCBDC 2023, August 17–19, 2023, Manchester, United Kingdom
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0733-9/23/08.
https://doi.org/10.1145/3616131.3616132

KEYWORDS

Data Complexity Measure, Instance Hardness Measures, Meta-Feature, Class Overlapping, Meta-Learning

ACM Reference Format:
Omaimah Saif Al Hosni* and Andrew Starkey. 2023. Investigating the Performance of Data Complexity & Instance Hardness Measures as A Meta-Feature in Overlapping Classes Problem. In *2023 7th International Conference on Cloud and Big Data Computing (ICCBDC 2023), August 17–19, 2023, Manchester, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3616131.3616132

1 INTRODUCTION

A common problem in meta-learning is establishing a (good) collection of meta-features that best represent the dataset properties [5]. In other words, the meta-learning recommendation's quality depends on the meta-features decision quality, and their ability to reflect the actual data challenges for the given dataset. Hence, the research question of this study is to what extent meta-features can describe the actual data difficulty without being affected by complex data challenges and thereby produced biased recommendation? According to literature, this question has not been given much attention but instead most of the works in this context focus on validating the meta-learning recommendation by evaluating the learning algorithms prediction performance (i.e., identifying correlations between meta-learning outputs and learning algorithm performance). From our study point of view, examining this correlation is not a good independent indicator to validate the complexity measure performance in estimating the actual data difficulty nor for showing the causes of the poor prediction of the learning algorithm's performance, since the complex data characteristics might also affect measures' performance in not reflecting the actual data difficulty and thereby produced biased meta-learning recommendation. In addition to that, both perspectives (learning algorithm and the Measures) adapt different assumptions and thereby react differently based on their sensitivity to the different data challenges. Accordingly, relying only on the learning algorithm performance to validate the measure performance might produce misleading information. Thus, in this work, the analysis of learning algorithm performance will be omitted from this study.

However, as the meta-feature analysis is data dependent, this study assumes that complex data characteristics such as class overlap might negatively affect the meta-feature decisions in not reflecting/estimating the actual difficulties (either by overestimating or underestimating the complexity), which in some cases would result in biased meta-learning recommendations that do not reflect the actual data properties. The reason for study to focus on the class overlap problem since this issue has long been recognised as one of classification's most challenging and pervasive problems. According to [6], 60–80% of overlapped samples are recognised as noise by the noise filters. Moreover, in real-world problems, datasets can have different geometrical class distributions and are usually associated with several issues, such as high-class overlapping caused by noisy features. In fact, several studies have concluded that class overlap, and difficult border decisions are significant contributors to degrading prediction accuracy [7], [8], [9], [10], [11]

Therefore, this study is interested in investigating this by using Complexity Measures (global data level) & Instance Hardness Measures (local/instance level) as a meta-feature to describe the data properties under varying data class overlap degrees. The motivation of using these measures (among different meta-feature methods proposed in the literature) is because these measures are mainly proposed to estimate the data complexity according to the geometrical descriptions of the shape and size of the decision boundaries, specifically on the class overlap imposed by feature values, separability, and data distribution in which match the data problem that this study interested to investigate.

The analysis will be conducted from both (global data level) and (local/instance level) measures perspectives to investigate to what extent these measures are able to describe the actual data difficulty without being affected by complex data challenges through addressing the following research questions:

- Are these measures able to reflect the actual data difficulty imposed by: High-class overlapping caused by noise features?
- Comparing global/local level measures, which one can best represent the actual difficulty of the complex data characteristics if any?

Answering the above questions can help to provide meaningful insights for the practitioners and researchers to choose the correct measures that are more appropriate for a particular dataset and give confidence in the output provided by that measure and improve the meta-learning recommendation at the end.

The paper is organized into five sections; Related Works section discusses the recent work that applied these measures in the context of meta-learning from both data-level (Data Complexity Measures) and instance-level perspectives, along with other works done to compare the performance of the measures from both perspectives. Methodology section represents the study methodology explaining the experimental setup designed for this work, followed by the experimental results in Results and Discussion section. Finally, the conclusion and future works will be presented in Conclusion section.

2 RELATED WORK

During the last decade, many meta-feature measures/methods have been proposed to describe the characteristics of the learning problems. Among these measures, there is a growing trend in the meta-learning field to use Data Complexity Measures as meta-features which act as descriptors of the spatial distribution of the data [1], [7]. Complexity Measures have been widely used as pre-processing steps to estimate the difficulty of the classification problem according to the topological characteristics of the dataset in separating the data points into their expected classes [8], [12]. The Complexity Measures analysis is conducted by extracting the geometrical descriptions of the shape and size of the decision boundaries, focusing on the class overlap imposed by feature values, separability, and data distribution [3], [13]. The popularity of these measures comes from the fact that several studies have concluded that class overlap, and difficult border decisions are significant contributors to degrading prediction accuracy [7], [8], [9], [10], [11].

It is worth noting that Complexity Measures (global data-level) were proposed by [14] and have a variety of uses for pre-processing data tasks, such as noise identification [15] and exploring the domain competence of different machine learning algorithms, which can help in hyperparameter optimisation [3]. Therefore, it is expected that these measures can play an essential role in improving the recommendation of the meta-learning model [13]. Further enhancement has been added to these measures by [15] to understand the difficulty of the imbalanced classification by considering each class individually including the minority class which was neglected by the original Complexity Measure as per [15] study.

In addition, Instance Hardness Measures (local/instance-level) have been proposed by [9] since the original data Complexity Measures focus on characterising the overall complexity of the entire dataset and fail to provide information at the local/instances level. The authors emphasise that obtaining such information is important to identify the misclassified instances and understand the causes that lead the instance to misclassify [9]. Instance Hardness Measures appear to have received less attention in the meta-learning context in literature compared to global Complexity Measures, which might be due to the latter's seniority.

In terms of applying global Complexity Measures in the Meta-learning context, a recent study conducted by [16] proposed a new taxonomy focused on imbalanced and overlapped classes issues and highlighted the current state of the Complexity Measures in meta-learning with other research areas. Another study by [17] established a meta-learning model using Complexity Measures as a meta-feature to predict the a priori performance of the Customized Naïve Associative Classifier (CNAC) and the measures have shown promising results in predicting (CNAC) performance. Furthermore, [18] proposed a novel meta-learning system to decrease the computational cost of data complexity measures while preserving their descriptive ability. With similar motivation, [19] have used a meta-learning approach to estimate the decomposed data Complexity Measures computing cost. Their study results indicated that the proposed approach is significantly faster yet more effective than computing the original Complexity Measures. On the other hand, [1] have provided an extensive list of meta-features and characterisation tools, including Complexity Measures, which can be used as a guide for the meta-learning practitioners.

As mentioned earlier, Instance Hardness Measures (instance/local-level measures) have received less attention in the context of meta-learning than global level measures. However, a recent interest has emerged in the research community applying Instance Hardness Measures in the meta-learning context. One of these studies [10] shows that the use of Instance Hardness Measures as a meta-feature allows for a deeper analysis of the relationship between instance hardness and classifier predictive performance by providing measures that vary according to the difficulty of each observation. Another study conducted by [20] used Instance Hardness Measures in meta-learning strategies to describe the key differences between easy-to-classify and hard-to-classify observations in a dataset. According to their study outcomes, this meta-knowledge can be used as a descriptor for characterising a data set's hardness profile and provided an insight into the leading causes of the difficulties they represent in the dataset. Furthermore, a comprehensive survey done by [7] aims to review the applications of both Complexity Measures and Instance Hardness Measures in different areas, including the meta-learning field. Moreover, to overcome the limitation of the global Complexity Measures in characterising the overall complexity of the entire dataset, [12] have decomposed some of the global measures into instance/ local-level measures, so the analysis is conducted based on the individual contribution of each instance instead of global complexity of the entire dataset from the class prospective. Then, they compared the performance of the proposed decomposed instance/local-level measures against the global equivalents and concluded that the former provided better performance than the latter.

Despite the advances shown in the recent work, an empirical comprehensive review of the ability of these measures (from both global and local perspectives) to give an estimation of the difficulty of a given data problem independent of the learning algorithm has not yet been undertaken. Most of the works undertaken in the literature are limited to examining the correlation between the values of the measures with the learning algorithms' prediction accuracy performance. From our study point of view, examining this correlation is not a good independent indicator to validate the complexity measure performance in estimating the actual data difficulty nor for showing the causes of the poor prediction of the learning algorithm's performance. Thus, in this work, the analysis of learning algorithm performance will be omitted from this study.

3 METHODOLOGY

As mentioned earlier, real-world problem datasets have a complex structure that usually suffers from several issues concurrently, such as high-class overlapping, high data sparsity, and complex decision boundaries. However, since most learning algorithms are data-dependent, knowing the causes of poor prediction accuracy is a nontrivial task, especially with the interactive effect of these data challenges, making it hard to identify the actual causes/data challenges that lead to poor performance of the learning algorithms. Thus, to better understand the actual effects of the different data challenges on measures performance, it is crucial to have a controlled environment that enables us to assess the effect of each data challenge on the measure's behaviour individually, and therefore synthetic datasets are used in this study. Furthermore, generating synthetic datasets will enable us to control the data difficulty by creating gradually increasing difficulty levels, starting from a manageable level, and moving to more challenging levels. The aim of using graded difficulty levels is to explore the interactive effect of different data challenges on the measures performance and to cover common real-world scenarios. Therefore, synthetic datasets are created at three levels of difficulty starting from the an easy level (Level One, linearly separable no overlap between the classes) to more challenging levels that are partially linearly separable (Level Two), and ending at Level Three where the classes are non-linearly separable as shown in Fig. 1. This will help us to investigate the measure's ability in reflecting the difficulty imposed by the class overlap and gives a description visually and also in terms of noise levels shown in Table 1 for the synthetic datasets constructed for the experimental results presented later.

To control the classes overlapping degree, Gaussian noise has been added to the classes' distribution according to the gradually increasing difficulty level, with the number of features remaining constant at 6. The characteristics of generated datasets is shown in Table 1 corresponding to each difficulty level.

As the objectives of this study is to conduct a comprehensive empirical study to compare the performance of Complexity Measures (the global level) and Instance Hardness Measures (instance local level), thus, the most widely used measures in the literature are selected in the study experimental setup as listed in Table 2 and are categorised based on the problems that they are designed to focus on. The chosen of these measures are based on their popularity with related to the global measures, and the increased performance in local measures for some cases, both categories were included in this study, resulting in 35 measures overall to be evaluated.

However, in order to be able to compare the outputs of all of these measures, and since measures values are bounded between 0 representing manageable datasets and 1 for the most complex problems, we have classified the measure values into the categories described in Table 3.

4 RESULTS AND DISCUSSION

The experiments results are presented and discussed in this section.

4.1 Experiment Results:

The study results are presented in the below subsections in line with the measure categories mentioned in Table 4 and thereby discussed separately for their advantages and disadvantages.

4.1.1 Measures of Overlapping of Individual Feature Values. These measures have been proposed to describe the data complexity caused by class overlap through evaluating the features' discriminative power in separating the classes [7]. However, as the study aims to compare the performance of the global and the local level measures, thus, the Table 4 below represents the study results of the global-level measures against the local-level. In general, the experiment outcomes presented in Table 4 show that both global- and local-levels of Overlapping Individual Features Values Measures can differentiate the gradual difficulty levels imposed by different degree of class overlaps as the measures have started with low values at Level One (where the classes are linearly separable) and described the dataset in very-easy category. Then , with increasing of the noise level at upper levels, the measures' values have increased reflecting the additional complexity caused by the expansion of the class-overlapped region.

However, despite the measure's ability to differentiate the complexity across the difficulty levels, the results showed considerable variation between the values of the global and local measures at this measure's category. Specifically, the Complexity Measures (global-level) at Level Two, where the linearity assumption for class

Table 1: Synthetic Datasets Characteristics for Scenario 1

Dataset Title	DifficultyLevel	Gaussian Noise Value	Sample Size	Feature Size	ClassRatio
Blobs_1	Level One	1	1000	6	Balanced
Blobs_2	Level Two	4			25:25:25:25
Blobs_3	Level Three	8			

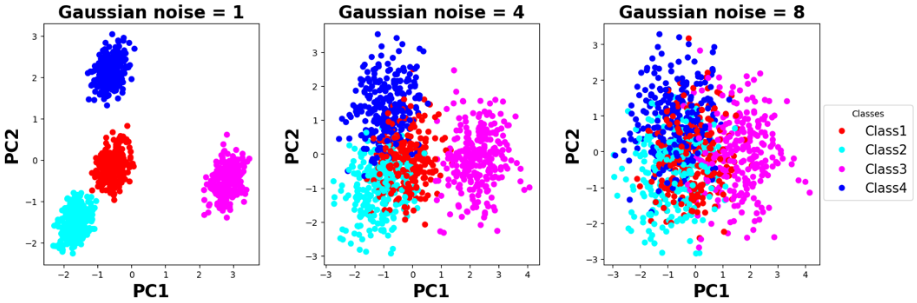

Figure 1: Dataset visualisation of Blobs Datasets with varying amounts of noise

separation still holds see Fig. 1 (considered as an easy problem), the global-level measures (F1, F2, F3 and F4) have produced high values categorising the problem in the range of medium to very difficult category (which is not the case) see Table 4 (Level Two column). The high values of the global-level measures at upper levels reflect that there is an overestimation of the problem's difficulty, given a relatively small amount of noise has been added at this level. According to [6], these measures are influenced by the number of features, the noise, and the sample size, which this study's results have also proven as the measures are heavily affected by the noise imposed in these levels Two and Three. Regarding the high value of F1 (the global level measure) and F1_HD (the local level measure), these measures estimate the difficulty using Fisher Linear Discriminant Analysis [8], which is also sensitive to noise [21].

In contrast, local-level measures (F2_HD, F3_HD, F4_HD) have provided a better representation of the problem at Levels Two and Three, categorising the complexity of the datasets in these levels as a medium category. However, as mentioned earlier, we excluded F1_HD, which performs similarly to F1 (the global level).

To sum up, in the category of Overlapping of Individual Feature Values Measures the study results indicate that despite the global-level measures (Complexity Measures) appear to perform better than the local measures in describing the underlying relationship of Level One, however local-level measure performed better in higher difficulty levels, specifically in F2_HD, F3_HD and F4_HD, albeit whilst still overestimating the complexity of the class overlap problem as the noise level increases.

4.1.2 Measures of Separability of Classes. The measures in this category are proposed to quantify the complexity according to the class separability by analysing the distance between the samples [22]. However, the study results in Table 4 show that all measures in this category can distinguish the complexity of the difficulty levels performing better than Overlapping of Individual Feature Values Measures in estimating the actual difficulty. In comparison between the global and the local-level measures of this measure's category, the results indicate that both global and local measures have shown very similar performance in describing the complexity across the difficulty levels. It is worth noting that N1, N2 and N3 (the global-level measures) aim to evaluate to what extent the classes are separable by examining the existence and shape of the class boundary. To estimate the complexity of the separability, N1 uses Minimum Spanning Tree (MST) method, while the remaining measures follow the nearest neighbour concept [23].

Regarding L1 and L2, the experiment results indicate that both measures can distinguish the gradually increasing difficulty levels and provide good estimation aligned with the difficulty levels. However, L1 has produced lower values than other measures in this category, classifying the dataset at Level Three within the easy range. The cause for such behaviour is that L1 does not check whether a linearly separable problem is more straightforward than another that is also linearly separable [7]. It is worth noting that both measures try to quantify to what extent the classes are linearly separable by creating a hyperplane to separate the classes using a linear classifier (usually SVM). For L1, it computes the sum of the distances of incorrectly classified samples to a linear boundary, whereas L2 computes the error rate of the linear SVM classifier. Therefore, both assume that a linearly separable problem can be considered simpler than a problem with a non-linear decision boundary [7].

4.1.3 Measures of Geometry, Topology and Density of Manifolds. The measures in this category try to capture the geometry of the manifolds covering each class by extracting information from the geometry (local) and topology (global) structure of the data to measure the class separability [8]. The results indicate that all measures can identify the simplicity of the datasets in Level One see Table 4. However, as the classes begin to overlap at the upper levels, the measures have responded differently. For example, in levels One and Two, N4 assigned almost identical values characterising the

Table 2: List of The Measures Used in This Study

Measure Categories		The Measures	Min	Max	Ref.
Overlapping of Individual Features Values	Global Level	F1: Maximum fisher's discriminant ratio	0	1	Ho &Basu (2002)
		F2: Volume of overlap region	0	1	
		F3: Maximum feature efficiency	0	1	
		F4: Collective feature efficiency	0	1	
	Local Level	$F1_{HD}$: Frac. feature values overlapping	0	1	Arruda et al. (2020)
		$F2_{HD}$: Volume of overlap region	0	1	
		$F3_{HD}$: Maximum feature efficiency	0	1	
		$F4_{HD}$: Collective feature efficiency	0	1	
Separability of Classes	Global Level	N1: Fraction of points on the class boundary	0	1	Ho &Basu (2002)
		N2: Ratio of inter/intra class nearest neighbour distance	0	1	
		N3: Leave-one-out error rate of the 1NN	0	1	
		L1: Sum of the error distance by linear programming	0	1	
		L2: Error rate of linear classifier	0	1	
	Local Level	$N1_{HD}$: Fraction of points on the class boundary	0	1	Arruda et al. (2020)
		$N2_{HD}$:Ratio of Inter/Intra class nearest neighbor distance	0	1	
Geometry, Topology and Density of Manifolds	Global Level	N4: Nonlinearity of a 1-NN classifier	0	1	Ho & Basu (2002)
		N5/T1: Fraction of hyperspheres covering data	0	1	
		L3: Nonlinearity of the linear classifier	0	1	
	Local Level	LSC: Local set cardinality	0	1	Leyva et al. (2015)
		LSR: Local set radius	0	1	
		H: Harmfulness	0	1	Leyva et al. (2014)
		U: Usefulness	0	1	
Data Sparsity & Dimensionality	Global Level	T2: Average number of features per points	0	1	Ho & Basu (2002)
		T3: Average number of PCA dimensions per points	0	1	Lorena, A.C. et al. (2012)
		T4: Ratio of the PCA dimension to the original dimension	0	1	
Structural Representation		Density: Average density of network	0	1	Garcia, et al., (2015)
		ClsCoef: Clustering coefficient	0	1	
		Hubs: Average hub score	0	1	
Instance Hardness Measures	Local Level	kDN : k-disagreeing neighbors	0	1	Smith et al. (2014)
		DS: Disjunct size	0	1	
		DCP: Disjunct class percentage	0	1	
		TDP: Tree depth pruned	0	1	
		TDU: Tree depth unpruned	0	1	
		CL: Class likelihood	0	1	
		CLD: Class likelihood difference	0	1	

Table 3: Measure Complexity Range

Complexity Category	Complexity Range
Very easy	0.00 - 0.10
Easy	0.11 – 0.30
Medium	0.31 – 0.50
Difficult	0.51 – 0.70
Very difficult	0.71– 1.00

dataset in both difficulty levels to be at the same complexity range which is very easy. The interpretation for such behaviour is that since this measure does not describe separability by design [24] and the classes at Level Two are still partially linearly separable, thus, it has assigned an identical value to Level One. Worth noting, this measure was initially proposed by [25] as a nonlinear measure to investigate the nonlinearity behaviour of pattern classifiers to a given data set by creating a new example interpolation of the training set. Then, it was used by [14] as one of the Neighbourhood Measures that aim to characterise the shapes of the manifolds spanned by each class.

However, N4 shares the same strategy as L3 in estimating the complexity, but instead of using a linear classifier (SVM) like L3, the measure uses a nonlinear classifier, usually 1-NN [14]. Concerning L3 performance, the results show that L3 provides a better representation in estimating the difficulty levels.

In contrast, the local-level measures, LSC, LSR, H and U share the same behaviour of Overlapping of Individual Features Measures

Table 4: Experiment Results

Measure Categories	Measures	Level One	Level Two	Level Three	
Measures of Overlapping of Individual Features Values	Global Level	F1	0.10	0.53	0.78
		F2	0.10	0.57	0.76
		F3	0.09	0.83	0.97
		F4	0.09	0.83	0.97
	Local Level	F1_HD	0.27	0.81	0.94
		F2_HD	0.15	0.3	0.36
		F3_HD	0.25	0.39	0.42
		F4_HD	0.37	0.45	0.47
Measures of Separability of Classes	Global Level	N1	0	0.23	0.61
		N2	0.11	0.39	0.48
		N3	0.00	0.15	0.47
		L1	0.00	0.13	0.29
		L2	0.00	0.23	0.53
	Local Level	N1_HD	0.00	0.15	0.46
		N2_HD	0.11	0.39	0.49
Measures of Geometry, Topology and Density of Manifolds	Global Level	N4	0.00	0.03	0.20
		N5/T1	0.00	0.45	0.843
		L3	0.00	0.22	0.50
	Local Level	LSC	0.00	0.92	0.99
		LSR	0.01	0.7	0.80
		H	0.00	0.00	0.00
		U	0.00	0.92	0.99
Measures of Data Sparsity and Dimensionality	Global Level	T2	0.00	0.00	0.00
		T3	0.00	0.00	0.00
		T4	0.5	1	1
Measures of Structural Representation	Global Level	Density	0.81	0.87	0.91
		CIsCoef	0.22	0.33	0.45
		Hubs	0.80	0.818	0.87
Instance Hardness Measures	Local Level	kDN	0.00	0.17	0.49
		DS	0.00	0.60	0.82
		DCP	0.00	0.13	0.47
		TDP	0.00	0.34	0.4
		TDU	0.00	0.49	0.5
		CL	0.00	0.13	0.45
		CLD	0.00	0.12	0.37

in overestimating the complexity in Levels Two and Three and describing the problem in both levels as difficult and very difficult. On the other hand, the H measure has produced identical values across different difficulty levels describing the complexity of the dataset in these levels as very easy. However, [26] have stated in their study that these measures are seriously affected by noise and class overlapping, which this study result has also proven.

N5/T1 has shown similar performance to the local-level measures of this category, particularly in Level Three, where it describes the dataset in this level as very difficult; however, it is better in differentiating the difficulty between Level Two and Three as it categorises the datasets in Level Two to be at the medium complexity range.

It is worth noting that LSC, LSR, H and U measures were proposed initially by [23] and have been used as instance local measure of the N5/T1 by [12] since they follow the same concept in characterising the data complexity through building hyperspheres centred at each instance and bounded by its nearest neighbour (instance from a different class). However, the main difference between the global- and local-levels measures is that the global measure (N5/T1) focuses only on the large hyperspheres that include samples from the same class. In contrast, the local measures estimate the difficulty according to the cluster from the same class and its nearest neighbour [8].

4.1.4 Measures of Data Sparsity and Dimensionality.

These measures have been proposed to measure the data sparsity caused by the high dimensionality, which resulted in difficulty to extract meaningful information because of the low-density areas imposed by data sparsity. Here we examine the performance of these measures in the class overlap problem; according to [7], this category has three global-level measures, which are T2, T3 and T4.

The experiment outcomes show that these measures cannot capture the complexity caused by the class overlap. The cause for such performance is that these measures estimate the data complexity according to the number of features the dataset has. Hence, since the datasets in this scenario include only six features (which are all relevant), T1 and T2 have produced low values across all different difficulty levels indicating that the problem is simple. It is worth noting that T2 is proposed by [14] which divides the number of samples in the dataset by their dimensionality. In contrast, T3 proposed by [27] estimates the complexity according to the data variability using Principal Component Analysis.

Regarding T4, this measure was proposed by [27], to provide an approximation of the ratio of the relevant features needed in the dataset by calculating the ratio of the PCA dimension to the original dimension. The larger T4 values mean more relevant features are needed to describe data variability in these datasets [7]. However, since the underlying relationship of the features at upper levels is affected by noise, T4 has overestimated the complexity across all difficulty levels describing the dataset at Level One as medium complexity and in Levels Two and Three as very difficult, reflecting that more relevant features are needed to describe the data variability due to small feature size (6 features) compared to the number of samples (1000 samples).

4.1.5 Measures of Structural Representation.

These measures have been proposed by [22] to characterise the data complexity according to the structural representation of the data set using graphs. This category has three measures, which are Density, ClsCoef and Hubs that applies transformation techniques to model the dataset into a graph in which each sample corresponds to a node/vertex [22]. However, to estimate the complexity, the relationship between the samples is modelled by preserving the similarity between sample pairs, in which their edges are connected and weighted by the distances between the samples [7].

The study outcomes indicate that Density and Hubs have assigned high values to all difficulty levels, denoting a very difficult problem, as shown in Table 4. In contrast, ClsCoef has a better representation of the class's overlapping problem as it can capture the gradually increasing difficulty levels.

4.1.6 Instance Hardness Measures.

As mentioned earlier, Instance Hardness Measures (local level)were proposed by [9] to overcome the limitation of the global Complexity Measures (global level) in which they estimate the data complexity at the sample level by identifying which samples are frequently misclassified in a dataset using various learning algorithms.

The experiment results show that all measures in this category can distinguish the gradual class overlapping degree across three difficulty levels Table 4 shows that kDN, DCP, CL and CLD perform better than other measures in this category in differentiating between Levels Two and Three. On the other hand, TDU characterises the datasets in Levels Two and Three at the same complexity range which is medium. The reason for the measures' inability to distinguish between both levels is that TDP and TDU estimate the difficulty by building a decision tree using the C4.5 Classifier which is well-known as being sensitive to noisy features [28]. However, as TDP is the pruned version, therefore, it is more robust to the noise than the unpruned version TDU, where the former has described the complexity better than the latter with TDU giving almost identical values in both levels. It is worth noting that both measures estimate the hardness of a sample by measuring the length of the tree depth, in which samples that are difficult to classify are typically placed at lower tree levels and have higher TD values [10].

Concerning DS and DCP, these methods adopt the disjunctive learning concept using the same C4.5 Classifier, which divides the task space to measure the overlap of samples; the measure DS is an unpruned method while DCP is a pruned method. In comparison, DCP is better at describing the actual difficulty than DS (for the same reason mentioned above). The result indicates that DS has overestimated the complexity in both levels Two and Three and classified the datasets at these levels to range from difficult to very difficult problem.

4.1.7 Summary.

The aim of this study is to investigate the performance of the measures under different degrees of noise which correspond to increasing levels of overlapping classes; the results indicate that most measures can correctly describe the underlying relationship of the relevant features in Level One. However, in Levels Two and Three, when the degree of noise increases, some measures have overestimated the complexity while a few have underestimated it, as shown in Table 5 which gives a summary of the performance of the metrics, identifying their level of performance across the three Levels of problem difficulty. In contrast, all Separability of Classes Measures, and Instance-Hardness Measures (apart from DS) show good performance in estimating the actual data challenges aligned with the difficulty levels.

5 CONCLUSION

In this study, we aimed to investigate to what extent meta-features can reflect the actual data difficulty without being affected by complex data challenges and produce values that either overestimate or underestimate data complexity. Highlighting this point is crucial to ensure the success of the meta-learning model since the meta-learning recommendation's quality depends on the meta-features decision quality. In real-world problems, the most common issues that are significant contributors to degrading prediction accuracy are high-class overlapping caused by noisy features. Accordingly, the Complexity Measures as global-level and Instance/local level measures are evaluated in this study as a meta-feature that act as descriptors of the spatial distribution of the data, for the problem of class overlapping caused by increased noise.

Since the meta-feature analysis is data-dependent, the study methodology has been designed to include the above data issue under gradually increasing difficulty levels starting from an easy problem and adding challenges at the upper levels. This has allowed the precise identification of how much the measure performance reflect the actual complexity and provides meaningful insights for the practitioners and researchers to choose the measures that are more appropriate for a particular dataset – and which ones not to use.

According to the research questions, the study results indicate that the measures responded differently to the above data issue. Some measures have overestimated the complexity while others

Table 5: Summary of the Measures' Performance Outcomes

Measure's Category	Good Estimate	Overestimate	Underestimate
Measures of Overlapping of Individual Features Values	F2_HD & F3_HD	F1, F2, F3, F4, F1_HD & F4_HD	-
Measures of Separability of Classes	N1, N2, N3, L1, L2, N1_HD & N2_HD	-	-
Measures of Geometry, Topology and Density of Manifolds	L3	N5/T1 , LSC , LSR , U	N4 & H
Measures of Data Sparsity and Dimensionality	-	T4	T2 & T3
Measures of Structural Representation	CIsCoef	Density & Hubs	-
Instance Hardness Measures	kDN , DCP, TDP , TDU , CL & CLD	DS	-

have underestimated it due to the different assumptions that the measures adapt and their sensitivity to the different data challenges. Based on the study result, and as described in the summary table, the measures that have provided good complexity estimation for the class overlap problem are F2_HD & F3_HD (the local-level measure from Overlapping of Individual Features Values Measures), all Separability of Classes Measures (global- & local-level), L3 from Geometry of Manifolds category, CIsCoef from Structural Representation Measures and all of the Instance Hardness Measures apart from DS.

This study has concluded that some measures do not perform well for the class overlapping problem and further work is required in order to determine how well the measures perform for different types of data problem such as data sparsity and more complex data relationships, and which previous studies have identified as affecting the complexity of a classification problem and the results obtained [7], [14].

Related to the research question in comparing the global- and local-level performance, the study outcomes have not provided a clear view as to which is the better approach. Future work will aim to investigate the measures' performance under different types of problems such as imbalanced classes, data sparsity and small sample sizes. A further conclusion to be drawn from this study is that Complexity Measures should be used with caution since they can be affected by different issues for the given dataset, which in the case of real-world data means they may easily under or overestimate the actual difficulty of the underlying data problem.

REFERENCES

[1] Rivolli, A., Garcia, L., Soares, C., Vanschoren, J. and de Carvalho, A., 2022. Meta-features for meta-learning. Knowledge-Based Systems, 240, p.108101.

[2] Tian, Y., Zhao, X. and Huang, W., 2022. Meta-learning approaches for learning-to-learn in deep learning: A survey. Neurocomputing, 494, pp.203-223.

[3] R. Shah, V. Khemani, M. Azarian, M. Pecht and Y. Su, "Analyzing Data Complexity Using Metafeatures for Classification Algorithm Selection," 2018 Prognostics and System Health Management Conference (PHM-Chongqing), 2018, pp. 1280-1284, doi: 10.1109/PHM-Chongqing.2018.00224.

[4] Garouani, M., Ahmad, A., Bouneffa, M., Hamlich, M., Bourguin, G. and Lewandowski, A., 2022. Using meta-learning for automated algorithms selection and configuration: an experimental framework for industrial big data. Journal of Big Data, 9(1).

[5] Lorena, A., Maciel, A., de Miranda, P., Costa, I. and Prudêncio, R., 2017. Data complexity meta-features for regression problems. Machine Learning, 107(1), pp.209-246.

[6] Gupta, S. and Gupta, A. (2018) "Handling class overlapping to detect noisy instances in classification," The Knowledge Engineering Review, 33. Available at: https://doi.org/10.1017/s0269888918000115.~

[7] Lorena, A., Garcia, L., Lehmann, J., Souto, M. and Ho, T., 2020. How Complex Is Your Classification Problem? ACM Computing Surveys, 52(5), pp.1-34.

[8] Barella, V., Garcia, L., de Souto, M., Lorena, A. and de Carvalho, A., 2021. Assessing the data complexity of imbalanced datasets. Information Sciences, 553, pp.83-109.

[9] Smith, M., Martinez, T. and Giraud-Carrier, C., 2013. An instance level analysis of data complexity. Machine Learning, 95(2), pp.225-256.

[10] Paiva, P., Moreno, C., Smith-Miles, K., Valeriano, M. and Lorena, A., 2022. Relating instance hardness to classification performance in a dataset: a visual approach. Machine Learning, 111(8), pp.3085-3123.

[11] Al Hosni, O. and Starkey, A., 2022. Assessing The Stability and Selection Performance of Feature Selection Methods Under Different Data Complexity. The International Arab Journal of Information Technology, 19(3A).

[12] Arruda, J.L.M., Prudêncio, R.B.C., Lorena, A.C. (2020). Measuring Instance Hardness Using Data Complexity Measures. In: Cerri, R., Prati, R.C. (eds) Intelligent Systems. BRACIS 2020. Lecture Notes in Computer Science, vol 12320. Springer, Cham. https://doi.org/10.1007/978-3-030-61380-8_33

[13] L. P. F. Garcia, A. C. Lorena, M. C. P. de Souto and T. K. Ho, "Classifier Recommendation Using Data Complexity Measures," 2018 24th International Conference on Pattern Recognition (ICPR), 2018, pp. 874-879, doi: 10.1109/ICPR.2018.8545110.

[14] Tin Kam Ho and M. Basu, "Complexity measures of supervised classification problems," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 289-300, March 2002, doi: 10.1109/34.990132.

[15] H. Barella, L. P. F. Garcia, M. P. de Souto, A. C. Lorena and A. de Carvalho, "Data Complexity Measures for Imbalanced Classification Tasks," 2018 International Joint Conference on Neural Networks (IJCNN), 2018, pp. 1-8, doi: 10.1109/IJCNN.2018.8489661.

[16] Santos, M.S. et al. (2023) "A unifying view of class overlap and imbalance: Key Concepts, multi-view panorama, and Open Avenues for Research," Information Fusion, 89, pp. 228–253. Available at: https://doi.org/10.1016/j.inffus.2022.08.017.~

[17] Tusell-Rey, C.C. et al. (2022) "A priori determining the performance of the customized naïve associative classifier for business data classification based on data complexity measures," Mathematics, 10(15), p. 2740. Available at: https://doi.org/10.3390/math10152740.~

[18] Garcia, L.P.F. et al. (2020) "Boosting meta-learning with simulated data complexity measures," Intelligent Data Analysis, 24(5), pp. 1011–1028. Available at: https://doi.org/10.3233/ida-194803.~

[19] Barella, V.H., Garcia, L.P.F., de Carvalho, A.C.P.L.F. (2020). Simulating Complexity Measures on Imbalanced Datasets. In: Cerri, R., Prati, R.C. (eds) Intelligent Systems. BRACIS 2020. Lecture Notes in Computer Science, vol 12320. Springer, Cham. https://doi.org/10.1007/978-3-030-61380-8_34

[20] Moreno, C.C. et al. (2021) "Contrasting the Profiles of Easy and Hard Observations in a Dataset," NeurIPS Data-Centric AI Workshop [Preprint].

[21] J. Wen et al., "Robust Sparse Linear Discriminant Analysis," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 2, pp. 390-403, Feb. 2019, doi: 10.1109/TCSVT.2018.2799214.

[22] Garcia, L.P.F., de Carvalho, A.C.P.L.F. and Lorena, A.C. (2015) "Effect of label noise in the complexity of classification problems," Neurocomputing, 160, pp. 108119. Available at: https://doi.org/10.1016/j.neucom.2014.10.085.~

[23] Leyva, E., Gonzalez, A. and Perez, R. (2015) "A set of complexity measures designed for applying meta-learning to instance selection," IEEE Transactions on Knowledge and Data Engineering, 27(2), pp. 354–367. Available at: https://doi.org/10.1109/tkde.2014.2327034.~

[24] Cano, J.-R. (2013) "Analysis of data complexity measures for classification," Expert Systems with Applications, 40(12), pp. 4820–4831. Available at: https://doi.org/10.1016/j.eswa.2013.02.025.~

[25] Hoekstra, A. and Duin, R.P.W. (1996) "On the nonlinearity of Pattern Classifiers," Proceedings of 13th International Conference on Pattern Recognition [Preprint]. Available at: https://doi.org/10.1109/icpr.1996.547429.~]

[26] Leyva, E., González, A. and Pérez, R. (2015) "Three new instance selection methods based on local sets: A comparative study with several approaches from a bi-objective perspective," Pattern Recognition, 48(4), pp. 1523–1537. Available at: https://doi.org/10.1016/j.patcog.2014.10.001.~

[27] Lorena, A.C. et al. (2012) "Analysis of complexity indices for classification problems: Cancer gene expression data," Neurocomputing, 75(1), pp. 33–42. Available at: https://doi.org/10.1016/j.neucom.2011.03.054.

[28] Mantas, C.J. and Abellán, J. (2014) "Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data," Expert Systems with Applications, 41(10), pp.4625–4637. Available at: https://doi.org/10.1016/j.eswa.2014.01.017.