Evaluation of oversampling data balancing techniques in the context of ordinal classification

Inês Domingues
José P. Amorim
IPO-Porto Research Centre
and CISUC, University of Coimbra
Email: icdomingues@dei.uc.pt

Pedro H. Abreu
CISUC, University of Coimbra
Coimbra, Portugal
Email: pha@dei.uc.pt

Hugo Duarte
and João Santos
IPO-Porto Research Centre
Porto, Portugal

***Abstract*—Data imbalance is characterized by a discrepancy in the number of examples per class of a dataset. This phenomenon is known to deteriorate the performance of classifiers, since they are less able to learn the characteristics of the less represented classes.**

For most imbalanced datasets, the application of sampling techniques improves the classifier's performance. For small datasets, oversampling has been shown to be the most appropriate strategy since it augments the original set of samples. Although several oversampling strategies have been proposed and tested over the years, the work has mostly focused on binary or multi-class tasks.

Motivated by medical applications, where there is often an order associated with the classes (increasing likelihood of malignancy, for instance), the present work tests some existing oversampling techniques in ordinal contexts. Moreover, four new oversampling techniques are proposed.

Experiments were made both on private and public datasets. Private datasets concern the assessment of response to treatment on oncologic diseases. The 15 public datasets were chosen since they are widely used in the literature. Results show that data balance techniques improve classification results on ordinal imbalanced datasets, even when these techniques are not specifically designed for ordinal problems. With our pipeline, better or equal to published results were obtained for 10 out of the 15 public datasets with improvements upon a decrease of 0.43 on MMAE.

I. Introduction

Class imbalance is a common problem in several domains, including both medical [1], [2], [3], [4], [5] and fraud detection [6], [7] settings. The detection of rare diseases is a typical example, where many samples without the disease are present while very few samples exist to represent the disease cases. When dealing with multi-class settings, this problem is typically even more present. While some attention has been given to the general multi-class case [8], [9], [10], its ordinal counterpart has not been given the same attention.

The most currently used strategies to deal with the data imbalance problem can be divided into data level (DAT) and algorithmic level (ALG) approaches [11]. While in the first, the objective is to preprocess the data so that the dataset becomes more balanced, the second is focused on developing classification algorithms specialized in handling this type of data.

Several ALG and DAT strategies to address the data imbalance problem have been proposed and evaluated for binary and (to a lesser degree) multi-class settings. The particular multi-class ordinal problem has not been so well studied by the community.

In this work, our main goal is to use non-ordinal data balancing techniques to address the ordinal data imbalance problem. While some algorithmic level approaches exist, we hypothesis that state of the art results can be achieved with data pre-processing. The main advantage of this approach is that typical supervised strategies can then be used for classification. Our work is targeted towards the detection and classification of less common diseases [2] (such as Hodgkin's Lymphoma and neuroendocrine tumors) where very large annotated datasets do not typically exist. To this end, we focus on oversampling as a means to augment the number of examples of less represented classes.

In this paper, a comprehensive literature review of works dealing with ordinal data-imbalance problems is performed. Four new data balance methods are also proposed and compared with state of the art techniques. A study on the data processing pipeline, in particular, an assessment on whether feature selection should be performed before or after data balancing, is also made.

Our results indicate that generic data balancing techniques have a positive effect on classification results in ordinal contexts. Moreover, we recommend feature selection before applying data balance techniques. With our pipeline, better or equal to published results were obtained for 10 of the 15 public datasets.

Next, we start by reviewing relevant works that deal with the ordinal data imbalance problem (Section II). Then, some techniques to deal with the data imbalance problem are described, both existing in the literature (Section III-A) and here proposed (Section III-B). The experimental setup is described in Section IV and results are given in Section V. Final conclusions and remarks are gathered in Section VI.

II. Literature Review

A dataset is imbalanced if there is one or more under-represented class(es) - the minority class(es) - in comparison with the other class(es) - the majority class(es). Imbalanced data is known to deteriorate the performance of classifiers [4].

978-1-5090-6014-6/18/$31.00 ©2018 IEEE

Ordinal classification (or regression) problems are those where the objective is to classify patterns using a categorical scale which shows a natural order between the labels [12]. Although ordinal problems tend to be imbalanced, with classes that are naturally more probable than others (especially when the number of classes is high) [13], there are not many works that tackle these problems simultaneously. Some notable exceptions are given next, where the acronyms DAT and ALG identify data and algorithmic level approaches, respectively.

The group led by Hervás-Martínez [14], [13], [3], [5] has made most of the contributions on this field, although other groups have also been working in the subject [15], [16], [17].

On the basic research side, an ordinal over-sampling method is developed in [13] (DAT). The method includes ordinal information by approaching over-sampling from a graph-based perspective. The results show a good synergy between the proposed method and support vector ordinal regression. A cost-sensitive version of the ordinal regression method is also introduced and compared with the over-sampling proposals showing, in general, lower performance for minority classes.

The main application of interest of this team is the allocation of organs for liver transplants [14], [3], [5]. The paper [14] (ALG) proposes a donor-recipient liver allocation system constructed to predict graft survival after transplant using the ordinal regression learning paradigm, via Error Correcting output codes with the utility cascade model and Support vector machines (ECSVM). Competitiveness is shown in all the selected metrics (designed for imbalanced and ordinal classification), when compared to other machine learning techniques and efficiently complements the Model for End-stage Liver Disease (MELD) score based on the principles of efficiency and equity. Subsequently, in [3] (DAT + ALG), an artificial neural network model applied to ordinal classification was used for the same problem, combining evolutionary and gradient-descent algorithms to optimize its parameters, together with an ordinal over-sampling technique. The authors named the methodology as IM-ORNET (IMbalanced Ordinal neural NETwork). The evolutionary algorithm applies a modified fitness function able to deal with the ordinal imbalanced nature of the dataset. The results of the proposed method improve upon those attained in with ECSVM. Next, the team proposes to improve IM-ORNET by treating the class weights dynamically [5] (ALG). The results obtained by DIM-ORNET improve those obtained in previous studies (ECSVM and IM-ORNET).

Doyle *et. al.* [15] (ALG) propose an approach to predict disease progression in Alzheimers disease (AD) multivariate ordinal regression which inherently models the ordered nature of brain atrophy spanning normal aging (CTL) to mild cognitive impairment (MCI) to AD. To account for imbalanced subject numbers per class, the probabilistic predictions for each test case were recalibrated whereby the prediction per class was divided by the proportion of that class represented in the training set. The probabilistic predictions per test case and across all four classes were then re-normalised to sum to one. The method was applied to 1023 baseline structural MRI scans. Distinguishing CTL-like from AD-like resulted in balanced accuracies between 79% and 82%. For prediction of conversion from MCI to AD, balanced accuracies of $70-15\%$ (AUC of $0.75-0.81$) were obtained.

Cruz *et al.* [16] (ALG) propose pair-wise ranking as a method for imbalance classification so that learning compares pairs of observations from each class, and therefore both contribute equally to the decision boundary. A new mapping threshold between ranking scores and ordinal classes is given, and the imbalance between pairs of differences is addressed. The same authors, in [17] extend ordinal classification using traditional balancing methods, such as pre-processing, training with costs, post-processing and ensembles. They verified that the application of weights is generally superior.

Achieved results in public datasets [18] are summarized in Table I. It is, however, necessary to carefully choose a measure to evaluate these types of problems. An imbalanced dataset has the consequence that, when testing a system with an evaluation measure conceived for balanced datasets, the trivial system assigning all items to a single class (typically, the majority class) may seem to outperform more advanced systems [19]. With the exception of MMAE, none of the metrics in Tables I are appropriate in an unbalanced setting.

It is clear from Table I that some datasets are harder to classify than others. The datasets *newthyroid*, *car* and *balance-scale* present, in general, good results, while *bondrate*, *ERA*, *SWD* have lower performances. This is likely due to the quality of the data itself, namely the existence of noise both at feature and class levels and the relevance of the features to the tasks.

As can be seen in the above literature review, most works develop algorithmic level approaches (five for ALG methods, while only two propose DAT techniques). While having the advantage of better representing the problem at hand, ALG methods have the disadvantage of not being readily available. DAT techniques are easier to implement and can be used in combination with well-established supervised learning strategies. On this basis, this is the approach followed in the present work.

III. Data balancing techniques

Different techniques were used to attack the data imbalance problem. The ones from the literature are summarized in Section III-A, while the ones proposed here are described in Section III-B.

A. Selected existing methods

From the existing techniques, five were considered relevant. Random oversampling and its noisy counterpart were chosen for their simplicity. SMOTE [6] and CBO [7] were selected for being two of the mostly commonly used methods, proving their relevance. Finally, DEAGO [23], an autoencoder based technique, was chosen for being a recent method that seems to present a promising approach to the problem. These are briefly described next.

TABLE I: Results presented in the literature (better results correspond to lower values). MAE = Mean Absolute Error; MMAE = Maximum MAE; MZE = Mean Zero-One Error.

Dataset	MAE	MMAE	MZE
automobile	0.35 [12]	0.783 [18]	0.246 [12]
	0.362 [18]	0.795 [13]	0.253 [18]
balance-scale	0.00 [20], [12]	0.103 [13]	0.001 [12]
	0.048 [13]	0.13 [17], [16], [21]	0.034 [22]
bondrate	0.53 [12]	0.2150 [18]	0.422 [22]
	0.576 [13]	1.633 [13]	[18], [12]
car	0.00 [12]	0.101 [13]	0.003 [12]
	0.012 [13]	0.27 [17], [16], [21]	
contact-lenses	0.33 [17], [21]	0.53 [17], [21]	0.244 [18]
	0.367 [18]	0.872 [18]	0.261 [12]
ERA	1.18 [12]	1.96 [16]	0.708 [12]
	1.21 [16]	2.134 [18]	0.738 [18]
ESL	0.297 [13]	0.95 [16]	0.284 [12]
	0.30 [16], [12]	1.026 [13]	0.316 [18]
LEV	0.41 [16]	0.91 [16]	0.371 [18], [22]
	0.40 [20], [12]	1.192 [18]	0.367 [12]
newthyroid	0.02 [12]	0.08 [16]	0.026 [12]
	0.025 [13]	0.09 [17], [21]	0.028 [22]
pasture	0.24 [12]	0.500 [18]	0.237 [22], [12]
	0.248 [18]		0.248 [18]
squash-stored	0.35 [12]	0.66 [17], [21]	0.359 [12]
	0.38 [17], [21]	0.813 [18]	0.398 [18]
squash-unstored	0.22 [12]	0.46 [17], [21]	0.221 [12]
	0.239 [18]	0.561 [18]	0.226 [18]
SWD	0.44 [20], [16], [12]	0.686 [13]	0.422 [12]
	0.442 [18]	0.69 [16]	0.426 [18]
tae	0.46 [12]	0.740 [18]	0.396 [12]
	0.523 [18]		0.434 [18]
toy	0.02 [12]	0.10 [17], [21]	0.020 [12]
	0.03 [17], [21]	0.137 [13]	0.147 [18]

1) Random oversampling: Here, a datapoint of the minority class is randomly selected and copied (hence, this technique will be named in the remaining of the document as "**Copy**"). As no variability is added to the dataset, this approach has been claimed to cause overfitting [2], the next strategies try to overcome this issue;

2) Random oversampling with noise: This technique differs from the previous one only in the fact that random noise is added to the copy of the selected datapoint (hence, this technique will be named in the remaining of the document as "**Copy + Noise**"). In our implementation, a maximum allowed noise of 0.1 times the feature standard deviation is allowed;

3) **SMOTE***: Synthetic Minority Over-sampling TEchnique:* In the SMOTE algorithm, the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors [6]. Results presented in this paper are for the case of three nearest neighbors;

4) **CBO***: Cluster-Based Over-sampling:* CBO [7] consists of clustering the training data of each class separately with the k-means technique and then performing over-sampling in each cluster. Here, over-sampling is performed both randomly and with SMOTE. Concerning the clustering evaluation criterion, Calinski Harabasz (CH) [24], Davies Bouldin (DB) [25], and silhouette [26] were used;

5) **DEAGO***: DEnoising Autoencoder-based Generative Oversampling:* An autoencoder consists of an encoding process and a decoding process. In the encoding process, the autoencoder learns a set of encoding weights to construct a code vector given the input vector. In the decoding process, it learns another set of decoding weights to map the code vector into an approximate reconstruction for the input vector [27].

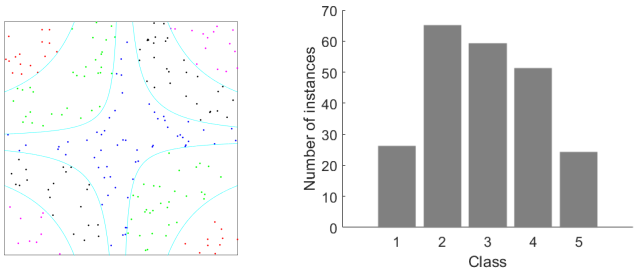
Here, new samples of the less represented class are given by the reconstruction of randomly selected noisy input samples of the less represented class [23].

B. Proposed methods

Four new methodologies are proposed in this work. Their details are illustrated with the artificially generated toy data in Figure 1a corresponding to the first fold of the dataset "toy" in [18] (and to the "toy" dataset of the experiential section IV-A). The unbalanced distribution of the random variable is shown is Figure 1b.

*1) Two become three (***2to3***):* Here, two points of the minority class are randomly selected and a new point is generated by taking their mean. Figure 2b clearly shows the problem of this technique when dealing with classes that are composed of clusters. The problem is particularly relevant when a cluster of points of a different class lies between the two clusters of the under-represented class. Here, new points can be generated in a subspace belonging to a different class [1];

[1]A possible way to deal with this problem would be to cluster the data in each class and generate new samples with points from the same cluster. The difficulty would lie in the choice of the number of clusters.



(a) First fold of "toy" dataset. (b) Class distribution

Fig. 1: Toy data used to illustrate the data balancing techniques. On the left, Class 1 is given in red, 2 in green, 3 in blue, 4 in black and 5 in magenta; theoretical boundaries are shown in cyan.

2) Feature by feature: Here, a point of the minority class is first selected. Then, for each feature, the average or median value of the k-closest points is taken. Figure 3 illustrates the approach. Looking at Figure 2c, it can be seen that in this technique it is less likely that a new point is generated in a subspace belonging to a different class, as observed for the 2to3 technique;

3) Centroid based: In the Centroid based approach the examples of the minority class are first aggregated into k clusters. Next, the k centroids of the generated clusters are added to the dataset. By using k-means to generate c centroids, where c is the difference between the number of the datapoints in the current class and the class with the most examples, the results on Figure 2d are achieved;

4) PCA based: The Principal Component Analysis (PCA) based approach starts by computing the principal components (PCs) of the minority class dataset. When reconstructing a randomly selected datapoint, n of its PCs are randomly selected and their weight is replaced by zero. The number of PCs to eliminate, n, is randomly selected, with an upper limit of 50%.

As the running example contains only two features, and correspondingly two PCs, eliminating half of it may have a significant impact in the generated points, placing them in a subspace of a different class, Figure 2e.

C. Dependence on feature selection

As stated in the introductory section, Section I, one of the goals of the present work is to study the interaction between feature selection and data balance. Table II shows if the order between the application of the data balance and feature selection techniques is independent.

IV. Experimental set-up

This section details the experimental set-up, describing the datased used, the evaluation measures and other methodological details.

A. Datasets

Two different sources of data were used, the first is private and the second comes from the literature.

TABLE II: Data balance techniques comparison. Proposed methods are given in bold.

Technique	Independent of feature selection method?
Copy	Yes
Copy + Noise	Yes
2to3	Yes
Feature by feature	Yes
SMOTE	No
CBO	No
DEAGO	No
Centroid	No
PCA based	No

The private datasets origin from [28], [29], [30] and have been described in [31]. They correspond to parameters retrieved from PET/CT of patients with either Hodgkin Lymphoma or neuroendocrine tumors. CD12 and CD1 contain clinical parameters while ID12 and ID1 have features automatically extracted from the PET/CT. CD12 and ID12 have information from both pre and post treatment, while CD1 and ID1 have only pre-treatment information. For these datasets, results are obtained by leave-one-out.

The datasets from the literature are listed in [18] [2]. Each one of these datasets is composed of a partition of 30 sets of training examples and 30 sets of associated test examples which have been obtained using a holdout stratified technique using 75% of the patterns for training and the remaining 25% for testing.

Some characteristics of the datasets are given in Table III that includes the average Imbalance Ratio (IR)[3] measure given by

$$IR = \frac{1}{Q}\sum_{q=1}^{Q} IR_q \tag{1}$$

where

$$IR_q = \frac{\sum_{j \neq q} N_j}{(Q-1)N_q} \tag{2}$$

and Q is the number of classes and N_q is the number of samples in the q^{th} class. The IR of a perfectly balanced dataset is equal to 1. Presented results correspond to average values of IR over all folds.

B. Evaluation measures

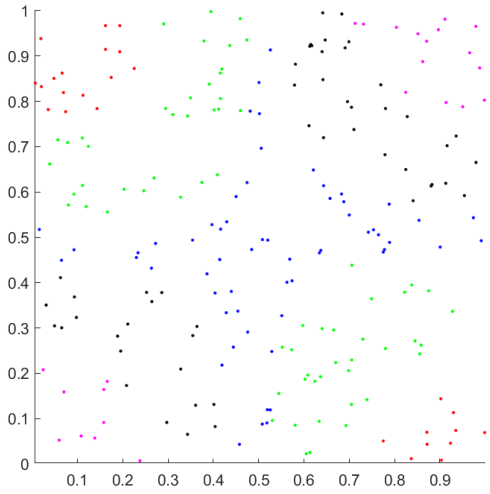
Evaluation measures for multi-class imbalanced domains are thorough discussed in [32]. There, it has been demonstrated that no single available measure is able to properly represent the results in this setting. When it comes to ordinal imbalanced domains, the consensus is even lower.

In the context of ordinal classification, the Mean Absolute Error (MAE) is widely used [21]. MAE is the average absolute deviation of the predicted class from the true class:

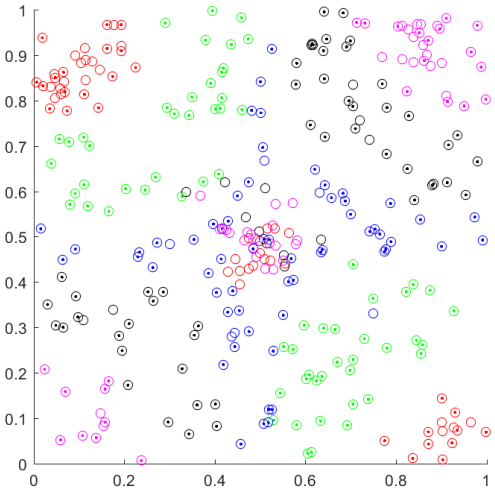
$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - y_i^*|$$

[2]datasets *eucalyptus*, *winequality-red*, *winequality-white*, *marketing* and *thyroid* were excluded due to time constraints.

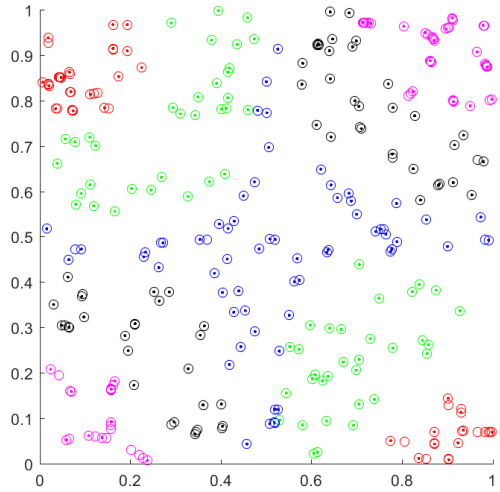
[3]adapted from [13]



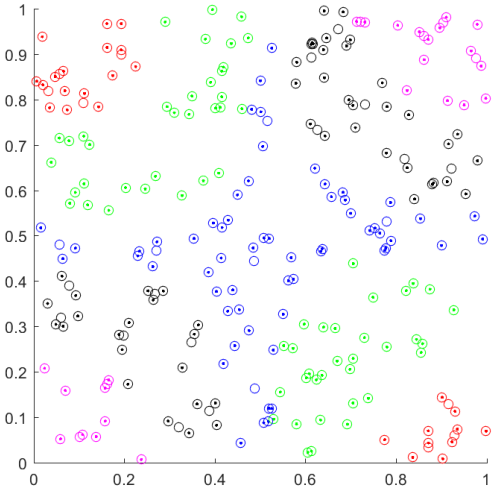
(a) Original dataset



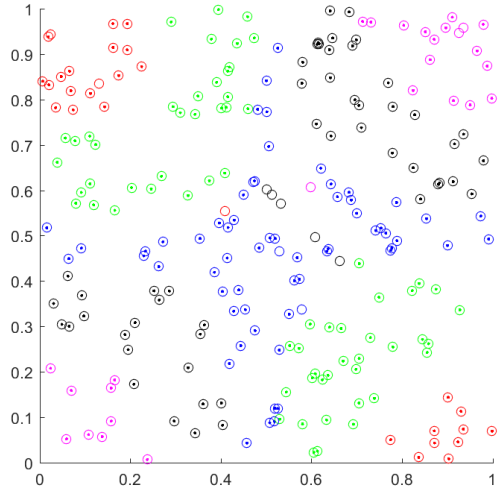
(b) 2to3



(c) Feature by feature



(d) Centroid based



(e) PCA based

Fig. 2: Illustration of the proposed methods on the first fold of "toy" dataset. Class 1 is given in red, 2 in green, 3 in blue, 4 in black and 5 in magenta. Dots correspond to the original data, circles represent the oversampled set. Data balance techniques: (a) 2to3, (b) Feature by feature, (c) Centroid based and (d) PCA based.

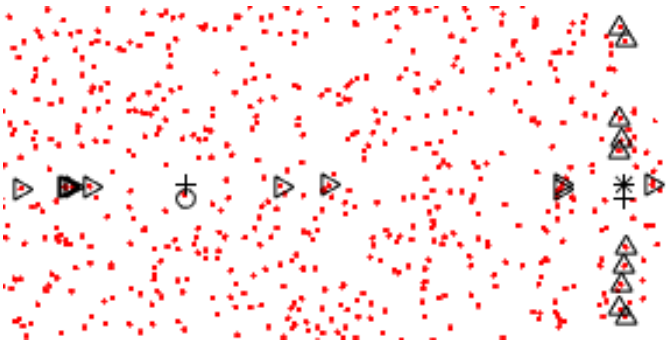


Fig. 3: Feature by feature data balance technique illustration. Dots in red represent the minority class. The selected point is illustrated by a black star. Nearest neighbors in the x-direction are given by black triangles facing up, while nearest neighbors in the y-direction are given by black triangles facing right. Median values for each dimension are illustrated by the plus sign. Finally, the generated point is given by a black circle.

TABLE III: Characteristics of the datasets

Dataset	# Features	# Classes	IR
CD12	20	3	1.58
ID12	383	3	1.58
CD12+ID12	403	3	1.58
CD1	11	3	1.58
ID1	147	3	1.58
CD1+ID1	158	3	1.58
automobile	25	6	3.33
balance-scale	4	3	2.39
bondrate	10	5	∞[4]
car	6	4	4.44
contact-lenses	4	3	1.57
ERA	4	9	1.85
ESL	4	9	5.75
LEV	4	5	2.65
newthyroid	5	3	1.94
pasture	22	3	1.00[5]
squash-stored	24	3	1.38
squash-unstored	23	3	2.39
SWD	10	4	3.10
tae	5	3	1.00
toy	2	5	1.24

where N is the total number of examples, y_i is the desired output for example i, and y_i^* is the prediction of the model.

MAE, however, does not take into consideration the imbalance nature of the data. Maximum MAE (MMAE) and Average MAE (AMAE) are two possible alternatives.

MMAE is the MAE value of the class with higher distance from the true values to the predicted ones [33]:

$$MMAE = max\{MAE_k, k = 1, \ldots, K\}$$

where MAE_k is the MAE for class k, and K is the total number of classes.

The AMAE is the mean of the MAE classification errors across classes [19]:

$$AMAE = \frac{1}{K}\sum_{k=1}^{K} MAE_k$$

All of these measures have values that range from 0 to $K-1$ [14].

C. Methodology

Feature selection, when performed, was made with Neighborhood Component Analysis (NCA) [34] or ReliefF [35]. These were selected for being filter methods (independent of the classification algorithm) and appropriate for multi-class problems. To evaluate if feature selection should be performed before or after data augmentation, both combinations were tested.

Concerning supervised learning methodologies, Discriminant analysis (DA) [36], Support Vector Machine (SVM) [36] and an ordinal scheme, Frank and Hall (F&H) [37] were used. F&H was instantiated both with DA and SVM. This selection was made based on previous work where this and several other techniques were compared [31]. Default Matlab R2016b (license 1056761) parameters were used.

Throughout we speak of two results as being "significantly different" if the difference is statistically significant at the 5% level (when not stated otherwise) according to a paired two sided t-test, where each pair of data points consists of the errors obtained in one of the runs of the two learning schemes being compared.

V. Results

From the results on the private datasets shown in Table IV, several observations can be made:

- Overall best results are obtained for CD1, while worst are for ID1. A similar observation had also been made in [31];
- To perform feature selection before applying data balancing is a better option;
- NCA is never the best feature selection technique;
- Although best overall results are obtained with DEAGO, CBO prevails as data balancing technique;
- F&H is most often the classifier present in the best model.

For the public datasets, it can be observed in Table V that:

- *bondrate*, *ERA* and *toy* datasets presented the worst results, while *balance-scale*, *car* and *newthyroid* achieved lowest errors. With the exception of the *toy* dataset with worse behaviour than *SWD*, these observations are in line with those published in the literature and summarized at the end of Section II[6];
- To perform feature selection before applying data balancing, is a better option;
- NCA is never the best feature selection technique;
- Not to perform data balance is the most prevalent setting;

[4]In 27 out of the 30 folds, the training set does not contain any example of class 5.

[5]All the training folds are perfectly balanced.

[6]As can be seen in Figure 1a, the toy dataset is very peculiar in the sense that not all of samples of one given class are "together". We hypothesize that this dataset will particularly benefit from methods developed specially for the ordinal case, so as to avoid the situation illustrated in Figure 2b.

TABLE IV: Results on private datasets. FS stands for feature selection. The symbol - under the "order" column means that the balance technique is independent of feature selection and thus the order does not change the results. Best overall results are given in bold. Statistically worse results are marked with *.

dataset	order	FS	oversampling	classifier	MMAE	AMAE
CD12	-	none	CopyNoise	F&H-SVM	0.30	0.18
CD12	SelThenBal	none	CBO SMOTE CH	F&H-SVM	0.30	0.14
ID12	SelThenBal	ReliefF	SMOTE	F&H-DA	0.30	0.18
ID12	-	ReliefF	Feature by feature	NB	0.33	0.17
CD12+ID12	-	none	Copy	F&H-DA	0.32	0.20
CD12+ID12	SelThenBal	ReliefF	CBO Rand DB	SVM	0.35	0.17
CD1	SelThenBal	ReliefF	DEAGO	DA	**0.25**	**0.13**
ID1	-	ReliefF	none	F&H-SVM	0.48*	0.30*
ID1	SelThenBal	none	CBO Rand CH	F&H-DA	0.49*	0.27*
CD1+ID1	SelThenBal	ReliefF	CBO SMOTE DB	NB	0.32	0.22
CD1+ID1	-	ReliefF	2to3	SVM	0.37	0.20

- SVM is most often the classifier present in the best model.

VI. Conclusions

This paper has extensively tested several data balancing techniques in the context of ordinal classification, both existing in the literature and four new proposed techniques. Our main conclusions include:

- Although specific methods tailored to the data imbalance problem in ordinal problems do exist, generic data balancing techniques can be used with the advantage of being readily available and easier to include in a data processing pipeline;
- Feature selection should be performed before data balancing;
- Neighborhood Component Analysis should not be used as a feature selection technique for ordinal imbalance datasets.

As a next step, we are interested in studying the relationship between the characteristics of the datasets (size, number of features, number of classes, imbalance ratio, etc.) and the models.

Although algorithm level approaches do seem to outperform data level ones in this context, we believe that there is still room for development of data level techniques for the ordinal imbalance problems, which have the advantage of modularity.

Another line of future work is to test and develop feature selection techniques tailored for ordinal imbalanced data, since it is clear from the present work that the used techniques are not adequate for this particular setting.

Acknowledgment

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).

References

[1] S. Bessa, I. Domingues, J. S. Cardosos, P. Passarinho, P. Cardoso, V. Rodrigues, and F. Lage, "Normal breast identification in screening mammography: a study on 18 000 images," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2014, pp. 325–330.

[2] M. S. Santos, P. H. Abreu, P. J. García-Laencina, A. Simão, and A. Carvalho, "A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients," *Journal of Biomedical Informatics*, vol. 58, pp. 49–59, 2015.

[3] M. Dorado-Moreno, M. Pérez-Ortiz, M. D. Ayllón-Terán, P. A. Gutiérrez, and C. Hervás-Martínez, "Ordinal Evolutionary Artificial Neural Networks for Solving an Imbalanced Liver Transplantation Problem," in *International Conference on Hybrid Artificial Intelligence Systems*, no. April, 2016, pp. 451–462.

[4] P. H. Abreu, M. S. Santos, M. H. Abreu, B. Andrade, and D. C. Silva, "Predicting Breast Cancer Recurrence Using Machine Learning Techniques," *ACM Computing Surveys*, vol. 49, no. 3, pp. 1–40, 2016.

[5] M. Dorado-Moreno, M. Pérez-Ortiz, P. A. Gutiérrez, R. Ciria, J. Briceno, and C. Hervás-Martínez, "Dynamically weighted evolutionary ordinal neural network for solving an imbalanced liver transplantation problem," *Artificial Intelligence in Medicine*, vol. 77, no. March, pp. 1–11, 2017.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, 2002.

[7] T. Jo and N. Japkowicz, "Class imbalances versus small disjuncts," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, p. 40, 2004.

[8] S. Wang and X. Yao, "Multiclass Imbalance Problems : Analysis and Potential Solutions," *IEEE Transaction on Systems, Man, and Cybernetics, Part B*, vol. 42, no. 4, pp. 1119–1130, 2012.

[9] L. S. D. Mosley, "A balanced approach to the multi-class imbalance problem," PhD dissertation, Iowa State University, 2013.

[10] S. Wang, L. L. Minku, and X. Yao, "Dealing with Multiple Classes in Online Class Imbalance Learning," *International Joint Conference on Artificial Intelligence (IJCAI)*, no. 5, pp. 2118–2124, 2016.

[11] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.

[12] P. A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernández-Navarro, and C. Hervás-Martínez, "Ordinal regression methods: survey and experimental study," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 1, pp. 127–146, 2015.

[13] M. Perez-Ortiz, P. A. Gutierrez, C. Hervás-Martínez, and X. Yao, "Graph-Based Approaches for Over-Sampling in the Context of Ordinal Regression," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 5, pp. 1233–1245, 2015.

[14] M. Pérez-Ortiz, M. Cruz-Ramírez, M. D. Ayllón-Terán, N. Heaton, R. Ciria, and C. Hervás-Martínez, "An organ allocation system for liver transplantation based on ordinal regression," *Applied Soft Computing Journal*, vol. 14, no. January, pp. 88–98, 2014.

[15] O. M. Doyle, E. Westman, A. F. Marquand, P. Mecocci, B. Vellas, M. Tsolaki, I. Kłoszewska, H. Soininen, S. Lovestone, S. C. R. Williams, and A. Simmons, "Predicting Progression of Alzheimer's Disease Using Ordinal Regression," *PLoS ONE*, vol. 9, no. 8, pp. 1–10, 2014.

[16] R. Cruz, K. Fernandes, J. F. P. Costa, M. Pérez-Ortiz, and J. S. Cardoso, "Ordinal Class Imbalance with Ranking," in *Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)*. Springer, Cham, 2017, pp. 3–12.

[17] ——, "Combining Ranking with Traditional Methods for Ordinal Class

TABLE V: Results on public datasets. FS stands for feature selection. The symbol - under the "order" column means that the balance technique is independent of feature selection and thus the order does not change the results. Results better than previously published are highlighted in bold.

dataset	order	FS	oversampling	classifier	MMAE	AMAE
automobile	-	none	Copy	F&H-DA	0.93	0.46
balance-scale	-	none	none	SVM	0.14	0.10
bondrate	SelThenBal	ReliefF	SMOTE	DR-DA	1.78	0.90
car	SelThenBal	none	CBO SMOTE DB	SVM	0.28	0.15
contact-lenses	-	none	none	DA	**0.48**	0.16
ERA	-	ReliefF	2to3	F&H-DA	**1.78**	1.29
ERA	-	none	Copy + Noise	F&H-SVM	**1.86**	1.26
ESL	-	none	Copy + Noise	DR-DA	**0.92**	0.42
ESL	-	ReliefF	2to3	F&H-SVM	**0.95**	0.37
LEV	-	none	Feature by feature	DR-DA	**0.76**	0.56
LEV	SelThenBal	ReliefF	SMOTE	DR-DA	**0.80**	0.55
newthyroid	SelThenBal	ReliefF	CBO SMOTE DB	F&H-SVM	**0.03**	0.01
pasture	-	ReliefF	none	NB	**0.50**	0.27
squash-stored	SelThenBal	none	CBO Rand CH	F&H-DA	**0.60**	0.37
squash-unstored	-	ReliefF	none	SVM	**0.46**	0.20
SWD	-	none	Copy	F&H-DA	**0.62**	0.49
SWD	-	none	2to3	SVM	0.70	0.48
tae	-	ReliefF	Copy	SVM	**0.74**	0.56
tae	SelThenBal	none	CBO Rand silhouette	SVM	0.76	0.55
toy	-	ReliefF	Copy + Noise	SVM	1.26	0.82

Imbalance," in *International Work-Conference on Artificial Neural Networks (IWANN)*. Springer, Cham, 2017, pp. 538–548.

[18] J. C. Gámez, D. García, A. González, and R. Pérez, "Ordinal Classification based on the Sequential Covering Strategy," *International Journal of Approximate Reasoning*, vol. 76, no. May, pp. 96–110, 2016.

[19] S. Baccianella, A. Esuli, and F. Sebastiani, "Evaluation Measures for Ordinal Regression," in *Ninth International Conference on Intelligent Systems Design and Applications*. IEEE, 2009, pp. 283–287.

[20] P. F. B. Silva and J. S. Cardoso, "Differential Scorecards for Binary and Ordinal data," *Intelligent Data Analysis*, vol. 19, no. 6, pp. 1391–1408, 2015.

[21] R. Cruz, K. Fernandes, J. F. P. Costa, and J. S. Cardoso, "Constraining Type II Error: Building Intentionally Biased Classifiers," in *International Work-Conference on Artificial Neural Networks (IWANN)*. Springer, Cham, 2017, pp. 549–560.

[22] J. Sánchez-Monedero, P. A. Gutiérrez, and C. Hervás-Martínez, "Evolutionary ordinal extreme learning machine," in *International Conference on Hybrid Artificial Intelligence Systems*, 2013, pp. 500–509.

[23] C. Bellinger, N. Japkowicz, and C. Drummond, "Synthetic oversampling for advanced radioactive threat detection," in *IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, 2015, pp. 948–953.

[24] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics - Theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[25] D. L. Davies and D. W. Bouldin, "A cluster separation measure." *IEEE transactions on pattern analysis and machine intelligence*, vol. 1, no. 2, pp. 224–7, 1979.

[26] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[27] Y. Ke, "Deep Networks Based Energy Models for Object Recognition from Multimodality Images," PhD dissertation, University of Sydney, 2016.

[28] M. A. Nogueira, "Creating Evaluation Functions for Oncological Diseases based on PET/CT," MSc dissertation, Universidade de Coimbra, 2015.

[29] M. A. Nogueira, P. H. Abreu, P. Martins, P. Machado, H. Duarte, and J. Santos, "An artificial neural networks approach for assessment treatment response in oncological patients using PET/CT images," *BMC Medical Imaging*, vol. 17, no. 1, p. 13, 2017.

[30] M. J. M. Tavares, "Previsão da evolução de doença oncológica a partir da análise de imagens de PET scan," MSc dissertation, Universidade do Porto, 2017.

[31] I. Domingues, P. H. Abreu, H. Duarte, and J. Santos, "Assessment of treatment response in oncological patients using PET/CT images: comparison of state-of-the-art supervised learning algorithms," SUBMITTED.

[32] P. Branco, L. Torgo, and R. P. Ribeiro, "Relevance-based Evaluation Metrics for Multi-class Imbalanced Domains," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, no. 1, 2017, pp. 698–710.

[33] M. Cruz-Ramirez, C. Hervás-Martínez, J. Sanchez-Monedero, and P. A. Gutierrez, "Metrics to guide a multi-objective evolutionary algorithm for ordinal classification," *Neurocomputing*, vol. 135, pp. 21–31, 2014.

[34] W. Yang, K. Wang, and W. Zuo, "Neighborhood Component Feature Selection for High-Dimensional Data," *Journal of Computers*, vol. 7, no. 1, pp. 162–168, 2012.

[35] M. Robnik-Siknja and I. Kononeko, "Theoretical and empirical analysis of RelifF and RReliefF," *Machine Learning*, vol. 53, pp. 23–69, 2003.

[36] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY: Springer New York, 2009.

[37] E. Frank and M. Hall, "A Simple Approach to Ordinal Classification," in *European Conference on Machine Learning*, 2001, pp. 145–156.