Pattern Analysis and Applications (2018) 21:931–939
https://doi.org/10.1007/s10044-018-0705-4

ORIGINAL ARTICLE

Binary ranking for ordinal class imbalance

Ricardo Cruz[1,2] · Kelwin Fernandes[1,2] · Joaquim F. Pinto Costa[3] · María Pérez Ortiz[4] · Jaime S. Cardoso[1,2]

Received: 30 September 2017 / Accepted: 30 March 2018 / Published online: 7 April 2018

Abstract

Imbalanced classification has been extensively researched in the last years due to its prevalence in real-world datasets, ranging from very different topics such as health care or fraud detection. This literature has long been dominated by variations of the same family of solutions (e.g. mainly resampling and cost-sensitive learning). Recently, a new and promising way of tackling this problem has been introduced: learning with scoring pairwise ranking so that each pair of classes contribute in tandem to the decision boundary. In this sense, the paper addresses the problem of class imbalance in the context of ordinal regression, proposing two novel contributions: (a) approaching the imbalance by binary pairwise ranking using a well-known label decomposition ensemble, and (b) introducing a regularization into this ensemble so that parallel decision boundaries are favored. These are two independent contributions that synergize well. Our model is tested using linear Support Vector Machines and our results are compared against state-of-the-art models. Both approaches show promising performance in ordinal class imbalance, with an overall 15% improvement relative to the state-of-the-art, as evaluated by a balanced metric.

Keywords Ordinal classification · Ordinal regression · Class imbalance · Learning to rank · Ranking

1 Introduction

In certain classification domains, class imbalance is pervasive. That is, the class distribution is not uniform, in some cases in an extreme fashion. This is a common problem in medicine and health care where there is a wide dispersion of patients suffering from different disease severities; it is inherent in fraud and fault detection where the anomaly is rare; and in many other fields.

Over-populated classes (denoted as majority classes from now on) can exert undue influence in the decision boundary. A naive classifier would have high accuracy by focusing on the majority classes, but have zero discriminatory power. Special metrics have been designed to evaluate the classifier and try to ensure it is unbiased. Typically, this problem is addressed by

1. introducing costs in the loss function,
2. pre-processing steps, such as undersampling or oversampling to alter class priors,
3. post-processing to tweak the model probabilities,
4. ensembles combined with pre-processing to ensure each model is trained with balanced data and the entire dataset is still used in aggregate.

Previously, we have introduced a fifth major alternative family of solutions:

5. using pairwise scoring ranking, which is a family of models borrowed from the learning to rank literature [4].

Two points about **pairwise scoring** ranking models:

- In **pairwise** ranking, observations are trained in pairs, which means there is no imbalance during training within a binary context;

✉ Ricardo Cruz
rpcruz@inesctec.pt

[1] INESC TEC, Porto, Portugal

[2] Faculty of Engineering, University of Porto, Porto, Portugal

[3] Mathematics Department and CMUP, Faculty of Sciences, University of Porto, Porto, Portugal

[4] Computer Laboratory, University of Cambridge, Cambridge, UK

- In **scoring** ranking, predictions are produced individually in the form of a score, making it possible to use them for classification.

While initially proposed for class imbalance in a binary context [4], this family of solutions has been extended to an ordinal context [5, 6]. This latter extension was not completely satisfactory; while the results were good, there was still imbalance present between pairs of different classes, which was addressed recurring to a traditional technique. In this work, we present an alternative solution to ordinal classification which is throughout balanced. The original ordinal problem is reduced to binary problems through the usage of an ensemble in the vein of Frank and Hall (F&H) [8]. The binary problems are solved in a balanced manner using pairwise ranking, and each vote within the ensemble has the same weight, ensuring balanced optimization in each step of the process.

One issue with the method of F&H is that the decision boundaries are not parallel across the models which make part of the ensemble. This is often criticized within the ordinal literature because it means decision boundaries will intersect at some point, making the ordinal prediction highly unstable to small perturbation within the intersection region [3]. The second novelty of the paper revolves around applying regularization across the models of the ensemble to ensure, at the limit, that decision boundaries are parallel to each other. Alternatively, if enough data is available, lower regularization provides some flexibility to allow biasing each decision boundary to better model each class frontier.

In essence, two novel ideas are presented in the paper: (i) an alternative solution to class imbalance in ordinal classification, within the methodology of pairwise scoring rankers, and (ii) an extension of F&H to allow regularizing the decision boundaries to better model ordinal data.

2 State-of-the-art

A vast literature exists for ordinal classification, also referred to as ordinal regression. First, the problem can and has been treated as a regression problem where the continuous prediction is discretized as a post-processing step. In this section, the focus is in the SVM family of models.

SVM, as originally formulated, solves a binary problem whose weights $\mathbf{w}$ and intercept (or bias) b are found by minimizing the hinge loss function,

$$J(\mathbf{w}) = \sum_{i=1}^{N} \max\left(0, 1 - y_i(\mathbf{w} \cdot \mathbf{x}_i + b)\right) + \lambda \|\mathbf{w}\|^2,$$

where $y_i \in \{-1, +1\}$ is the binary class label of observation $\mathbf{x}_i$. Two groups of ordinal classification solutions exist:

(i) solving the ordinal problem by explicitly manipulating the loss function,
(ii) turning the ordinal problem into several binary problems.

Within the first group, we can find such models as SVOR [2], which adds $K-1$ biases for the decision boundary, one between every two consecutive ordinal classes,

$$J(\mathbf{w}) = \sum_{k=1}^{K-1} \sum_{i=1}^{N} \max\left(0, 1 - y'_i(\mathbf{w} \cdot \mathbf{x}_i + b_k)\right) + \lambda \|\mathbf{w}\|^2,$$

where $y'_i = 1$ if $y_i = k$ and -1 otherwise. This formulation is based on the adaptation which can be found in [11]. The weights $\mathbf{w}$ are shared between decision boundaries, making them parallel to each other. An adaptation which allows flexibility of the decision boundaries, without allowing intersections can be found in [14]. Intersection of decision boundaries is seen as problematic, since it makes the ordinal prediction volatile within the intersection region [3]. Consider Fig. 1, it is not obvious what class should be placed in the "?" decision space. Most classifiers solve this problem by ensuring parallelized decision boundaries.

Within the second group of ordinal classifiers, most solutions involve an ensemble, which are known as binary decomposition methods [9]. An exception is oSVM which finds the different biases b_k through a pre-processing step which expands the feature space and a post-processing step which transforms the coefficients of the new features into biases [1]. Another alternative is the scoring pairwise solution as described in the previous section. Typical ensembles are solutions also used in multi-class cases, such as One-vs-Rest which trains K models and One-vs-One which trains $K(K-1)$ models. These ensembles suffer from the intersection problem previously described. Also, they do not take ordinal metrics in consideration; these metrics punish more highly misclassification of more distant classes.

A popular ensemble solution is the one proposed by Frank and Hall (F&H) [8]. This ensemble reduces the ordinal problem into $K-1$ traditional binary classification problems. For every $i \in \{1, \ldots, K-1\}$, the i-th model is trained using classes $\{\mathscr{C}_1, \ldots, \mathscr{C}_i\}$ against $\{\mathscr{C}_{i+1}, \ldots, \mathscr{C}_K\}$. If we have four classes, then three models are produced using class (i) $\mathscr{C}_1$ against $\{\mathscr{C}_2, \mathscr{C}_3, \mathscr{C}_4\}$, (ii) $\{\mathscr{C}_1, \mathscr{C}_2\}$ against $\{\mathscr{C}_3, \mathscr{C}_4\}$ and (iii) $\{\mathscr{C}_1, \mathscr{C}_2, \mathscr{C}_3\}$ against $\mathscr{C}_4$. Each model $f_i(\mathbf{x})$ is trained

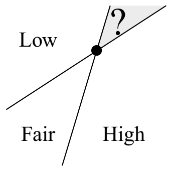

Fig. 1 Diagram showing the lexicographical violation when decision hyperplanes intersect within the feature domain

to produce 0 if $k \leq i$ or 1 if $k > i$. This makes each classifier use highly imbalance data, even if the data was not originally so. The final prediction is then a simple cumulative voting, $\hat{k} = 1 + \sum_{i=1}^{K-1} f_i(\mathbf{x})$.

While F&H takes the ordinal information into account, and therefore, produces better results as evaluated by ordinal metrics, the ensuing final model suffers from the intersection problem which violates the lexicographical order of the ordinal classes. That is, while voting for hypothetical class $\hat{k}$, model i may vote against $\hat{k} \geq i$ while model j, with $j > i$ may vote for $\hat{k} < j$.

3 Preliminary work

Four main approaches for addressing class imbalance can be identified from the literature: cost matrices, pre-processing, post-processing, and ensembles. Recently, a fifth proposal was introduced by the authors: pairwise scoring ranking. Initially, this solution was applied in a binary context [4], and then adapted to an ordinal context [5, 6].

3.1 Binary ranking

Several family of models exist within the learning to rank literature: pointwise, pairwise and listwise, to rank "documents", hereby termed "observations" [11]. In ranking, observation $\mathbf{x}_i$ is compared with another observation $\mathbf{x}_j$, and we are interested in predicting whether $\mathbf{x}_i \succ \mathbf{x}_j$, meaning $\mathbf{x}_i$ is "preferred" to $\mathbf{x}_j$.

The **pairwise** family of models is of particular interest to class imbalance because a function f is trained by comparing each observation $\mathbf{x}_i$ against all others $\mathbf{x}_j$. In the case of binary classification, with two classes C_- with N_- observations and C_+ with N_+ observations, pairwise rankers can be trained so that for every two observations (of different classes), $(\mathbf{x}_i, \mathbf{x}_j)$, and respective class labels, (k_i, k_j), f learns that $\mathbf{x}_i \succ \mathbf{x}_j$ if $\mathrm{P}(k_i = \mathscr{C}_i) > \mathrm{P}(k_j = \mathscr{C}_i)$, and $\mathbf{x}_j \succ \mathbf{x}_i$ otherwise. Here, $\mathscr{C}_i$ could be either one of the two classes, without loss of generality. Given that training between classes involves $2N_-N_+$ comparisons between all pairs of each class, $(\mathbf{x}_i, \mathbf{x}_j)$ with $k_i \neq k_j$, therefore, no class imbalance is present during training.

Within **non-scoring pairwise ranking**, function f decides which of two observations is preferred, $f : X \times X \rightarrow X$. In **scoring pairwise ranking**, this function is of the form $f : X \rightarrow \mathbb{B}$, and defined as $f(a, b) = s(a) > s(b)$ with $s : X \rightarrow \mathbb{R}$, so it is trained to produce a score such that $f(\mathbf{x}_i) > f(\mathbf{x}_j)$ when $\mathbf{x}_i \succ \mathbf{x}_j$. Therefore, we have proposed a framework whereby observations are trained in pairs, but predictions can still be performed individually. A threshold rule, $t : \mathbb{R} \rightarrow \mathscr{C}$, can then applied to convert the resulting ranking score into a class.

The threshold t can be chosen to maximize a balanced metric $m(k, \hat{k})$. This metric could be the F_1 score or G-mean, which are described in the Experiments section. Using the training data, we have $s_i = s(\mathbf{x}_i)$ which is sorted, and then each midpoint $s_i' = \frac{s_i + s_{i+1}}{2}$ is tested as possible candidates for threshold t, so that $t = \arg\max_{s_i'} m(k, \hat{k})$, where $\hat{k} = \mathscr{C}_-$ if $s_i < s_i'$ and $\hat{k} = \mathscr{C}_+$ otherwise.

A particular instantiation of pairwise training is to work on the space of differences [10]. A scoring pairwise ranker can be produced through a pre-processing step where the original data domain is transformed into a domain of differences

$$\{(\mathbf{x}_{ij}, +1), (-\mathbf{x}_{ij}, -1) \mid \mathbf{x}_{ij} = \mathbf{x}_i - \mathbf{x}_j\}, \quad (1)$$

where $+1$ and -1 are the labels k_{ij} and k_{ji} of the samples $\mathbf{x}_{ij}$ and $-\mathbf{x}_{ij}$, respectively. An illustration of this pre-processing step is shown in Fig. 2. When using a linear SVM estimator, without the bias term, the linear SVM is trained on the space of differences, and the decision rule $\mathbf{w} \cdot (\mathbf{x}_i - \mathbf{x}_j) > 0$ can then be transformed into a scoring function since $\mathbf{w} \cdot (\mathbf{x}_i - \mathbf{x}_j) > 0 \equiv \mathbf{w} \cdot \mathbf{x}_i > \mathbf{w} \cdot \mathbf{x}_j \equiv s(\mathbf{x}_i) > s(\mathbf{x}_j)$.

3.2 Ordinal ranking

Within the context of ordinal classification, we define the ranking lexicographical order to be the class order, $\mathbf{x}_i \succ \mathbf{x}_j$ if $i < j$ [6]. The reverse lexicographical order could, of course, also be chosen without affecting the results. The previous instantiation could be used to produce an ordinal pairwise scoring ranker. In fact, the ordinal data classification problem has already been addressed using pairs in the space of differences by Herbrich et al. [10].

However, class imbalance is not completely satisfied. Pairwise ranking involves taking every observation $\mathbf{x}_i^{(k)}$ belonging to class k, and training with all combination pairs $(\mathbf{x}_i^{(\ell)}, \mathbf{x}_j^{(m)})$ with $\ell \neq m$. In the binary context, it is trivial to see that every class is equally represented, since we can either have the pairs $(\mathbf{x}_i^{(\ell)}, \mathbf{x}_j^{(m)})$ or $(\mathbf{x}_i^{(m)}, \mathbf{x}_j^{(\ell)})$. Therefore, every class is contributing equally to the decision boundary. This is not the case in an ordinal context. Let N_k be the number of observations of class $\mathscr{C}_k$, then take, as an example, three classes $N_1 = 10$ and $N_2 = N_3 = 100$. Data from C_1 is contributing with 4,000 elements in the new

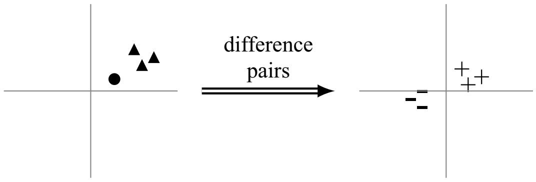

Fig. 2 Illustration of the differences space transformation

space, while the data from C_2 or C_3 is contributing with 22,000 each. So, the new learning problem will be dominated by the samples from C_2 and C_3 and it is likely that the decision boundaries C_1 will be poorly estimated. A solution would be to repeat pairs, but this could be computationally impractical.

In initial work, weights were used to overcome this imbalance problem [6], and then extended to other traditional techniques of class imbalance in [5]. The solutions presented improved current solutions, but did not *fully* balanced classes.

The threshold proposed in [6] is similar to that described in the previous section for the binary case. Except that testing all combinations would be $O(n!)$, as there would be C^N_{K-1} combinations to test. However, the fact that the ranking lexicographical order and the ordinal classes order are the same can be exploited; i.e. if $\ell < m$ then the model is trained so that $f(\mathbf{x}_i^{(\ell)}) < f(\mathbf{x}_j^{(m)})$ and therefore, $t_\ell < t_m$ must be true. Furthermore, the typical metric used in ordinal class imbalance contexts is the Mean Absolute Error (MAE), which is a summation, and therefore, can be computed incrementally, $\text{MAE} = \frac{1}{N}\sum_i |k_i - \hat{k}_i|$. Both of these conditions make it possible to reduce the search space and to compute the threshold in an incremental fashion. We have proposed doing so by minimizing the following recursive function, which can be used to generate a binary tree of possible thresholds and respective errors,

$$f(s_i, k_i, \hat{k}) = \begin{cases} 0 & \text{when } i = N \\ \varepsilon_{k_i\hat{k}} + f(s_{i+1}, k_{i+1}, \hat{k}) & \text{when } \hat{k} = K \\ \min\left\{ \varepsilon_{k_i\hat{k}} + f(s_{i+1}, k_{i+1}, \hat{k}),\ f(s_i, k_i, \hat{k}+1) \right\} & \text{otherwise} \end{cases}$$

$\varepsilon = \left[\varepsilon_{k\hat{k}}\right]$ is a cost matrix. Several cost schemes have been proposed: uniform, absolute and inverse class frequency matrices as costs. In the experiments section of this work, absolute costs have been used, $\varepsilon_{k\hat{k}} = \left\{|k - \hat{k}|,\ \forall k, \hat{k}\right\}$. Repeated evaluations of f can be avoided using dynamic programming.

4 Proposal

Two proposals are presented in this work. First, we introduce the notion of ranking into the Frank and Hall (F&H) ensemble to minimize the effect of class imbalance. Second, we propose a regularization scheme to solve the lexicographical order violation described in the previous section. While we advocate these two innovations be used in tandem, each stands on its own and does not require the other.

4.1 Ordinal classification as binary ranking

In previous work, we introduced pairwise scoring ranking for the binary case [4], and then extended for the ordinal case [5, 6]. This extension did not fully address class imbalance, as detailed in Sect. 3.2.

The proposal here described merges the binary scoring pairwise ranking from Sect. 3.1 with the F&H ensemble from Sect. 2.

The F&H ensemble is initially applied to convert the ordinal problem into $K-1$ binary problems, such that, during training, labels for each model f_i are given as

$$k' = \begin{cases} 0, & \text{if } k \leq i \\ 1, & \text{if } k > i. \end{cases}$$

Prediction is performed by voting whether the observation is to the left or to the right of the decision boundary, $\hat{k} = 1 + \sum_{i=1}^{K-1} f_i(\mathbf{x})$.

Several scoring pairwise rankers exist. In this work, we train a linear SVM in the space of differences. For all combinations of every two observations of the two different classes, the problem is solved in the set of the differences as shown in (1), except now +1 is the label of the observation $\mathbf{x}_{ij}$ when $k'_{ij} = 1$, and 0 is the label of the observation $-\mathbf{x}_{ij}$ when $k'_{ji} = 0$.

During prediction, the ranking score is transformed into a binary class using a threshold rule. The threshold is found during training by maximizing the F_1 metric as described in Sect. 3.1. The entire process is schematized in Fig. 3.

4.2 Regularizing Frank and Hall decision boundaries

F&H may violate the lexicographical order, as described in Sect. 2. To tackle this problem, we propose: (a) training a

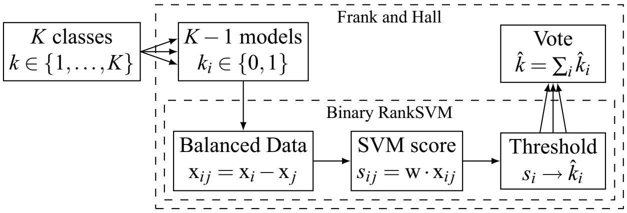
K classes
$k \in \{1, \ldots, K\}$
Frank and Hall
$K-1$ models
$k_i \in \{0, 1\}$
Vote
$\hat{k} = \sum_i \hat{k}_i$
Binary RankSVM
Balanced Data
$\mathrm{x}_{ij} = \mathrm{x}_i - \mathrm{x}_j$
SVM score
$s_{ij} = \mathrm{w} \cdot \mathrm{x}_{ij}$
Threshold
$s_i \rightarrow \hat{k}_i$

Fig. 3 Diagram showing the proposed balanced ordinal classification

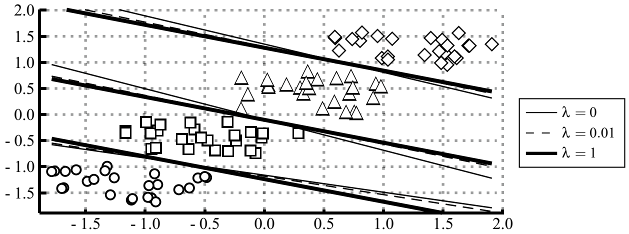

Fig. 4 Synthetically generated parallel classes

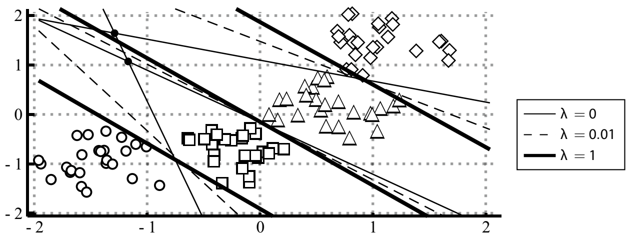

Fig. 5 Synthetically generated non-parallel classes. Black points represent hyperplane intersections

base ordinal estimator g, and then (b) regularize each f_i to not diverge much from g. This makes decision hyperplanes more parallel to each other, avoiding intersections within the feature domain.

Taking linear SVM as a case study, and depending on the model family, stage (a) of constructing a base estimator g can be performed using:

Ordinal classification: oSVM or SVOR (Sect. 2) may be used to estimate the weights $\mathbf{w}'$ of the model g,

Pairwise scoring ranking: the RankSVM model described in Sect. 3.2 may be used to estimate the weights $\mathbf{w}'$ of the model g.

For stage (b), the SVM formulation of each individual model must be tweaked so that regularization is related to the "base model" g,

$$J((w)) = \sum_{i=1}^{N} \max\left(0, 1 - y_i(\mathbf{w} \cdot \mathbf{x}_i + b)\right) + \lambda \|\mathbf{w} - \mathbf{w}'\|^2.$$

If regularization $\lambda = 0$, then the pure F&H solution is used. If regularization $\lambda \to \infty$, then the decision boundary orientation will be parallel as in the base ordinal estimator while only the bias is provided by f_i. Figures 4 and 5 reproduce this using synthetic data; $\lambda = 0$ result in flexible boundaries which intersect, violating class lexicographical order, $\lambda = 1$ result in approximately parallel boundaries, while $\lambda = 0.01$ offers an moderate approach.

This proposal can be seen as a simpler non-parallel alternative to solutions that involve overhauling the SVM loss function [14]. (Regarding bias, for each f_i, bias b may be initialized with the respective b_i from g to help convergence of the optimization algorithm.)

5 Experiments

The following models have been tested:

– SVOR, described in Sect. 2,
– ordinal RankSVM, see Sect. 3.2,
– the proposed binary decomposition using Frank and Hall (F&H), composed of two families of models:

  – SVM regularized against the base model SVOR,
  – binary RankSVM regularized against the base model ordinal RankSVM.

The implementation of the models and datasets can be found at https://github.com/rpmcruz/ranking-imbalance/tree/master/f-and-h. The models were written in Julia v0.5.

5.1 Synthetic data

Synthetic samples were generated as already illustrated in Figures 4 and 5. The data generated contains $K = 4$ classes with $N = 400$ in total, distributed in an imbalance fashion using a geometric series rule, $N_k = (\frac{1}{2})^k$.

The angle with which each class is generated is given by $a_k = (k - 1)\Delta a + a_1$, with both a_1 and Δa given as parameters. The difference between the aforementioned figures is only in Δa ($\Delta a = 0$ for Fig. 4, and $\Delta a = 30$ degrees for Fig. 5).

Furthermore, the noise is given by an uniform random generator and two ranges are parametrized; the noise with respect to angle a_k (ϵ) and the noise in the orthogonal direction (ϵ').

5.2 Real datasets

The datasets used comprise a subset of fifteen of the datasets in [12], as illustrated in Table 1. The selection was chosen based on dataset size, due to ranking being highly computationally demanding. For evaluation purposes, each dataset was divided in the same 30 folds as in [12].

The Imbalance Ratio (IR) is the arithmetic average of the per-class k imbalance ratios as given by:

$$\mathrm{IR}_k = \frac{\sum_{i \neq k} N_i}{(K - 1)N_k}.$$

Table 1 Characteristics of the datasets used in the experiments

Dataset	N	Features	K	**IR**	OR
toy	300	2	5	1.25	0.11
wisconsin5	194	32	5	1.27	0.74
wisconsin10	194	32	10	1.31	0.85
diabetes10	43	2	10	1.53	0.79
contact-lenses	24	6	3	1.57	0.33
ERA	1000	4	9	1.86	0.87
diabetes5	43	2	5	1.98	0.51
auto5	392	7	5	2.60	0.45
LEV	1000	4	5	2.70	0.69
auto10	392	7	10	2.72	0.62
SWD	1000	10	4	3.10	0.62
machine10	209	6	10	3.55	0.42
machine5	209	6	5	3.55	0.20
ESL12vs3vs456vs7vs89	488	4	5	3.66	0.22
ERA1vs23456vs7vs8vs9	1000	4	5	5.32	0.49

A dataset is balanced when IR = 1, and usually considered imbalanced when IR > 1.5. Table 1, as well as the following ones, are ordered by this imbalance ratio metric.

Another metric displayed is the Overlap Ratio (OR), first introduced in [4], and here extended for the first time to the ordinal case. The metric is defined as

$$\mathrm{OR} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left(y_i \neq y_j,\ j = \arg\min_{k}(\mathbf{x}_i - \mathbf{x}_k)^2\right),$$

i.e. the ratio of observations having an observation of another class as its nearest neighbor, using Euclidean distance. Data is normalized before this computation using the standard score. This metric was designed to test the hypothesis advanced by some authors that class imbalance problems are often due to overlapping problems [7].

5.3 Metrics

Typically, in binary imbalance problems, specially designed metrics are used. The most popular are F_1 and G-mean.

F_1 is defined as $\mathrm{F}_1 = \frac{2\,\mathrm{TP}}{2\,\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$, where TP, FP and FN are true positives, false positives, and false negatives, respectively. While, G-mean is the geometric average between sensitivity and specificity, G-mean $= \sqrt{\frac{TP}{TP+FN}\left(1 - \frac{FP}{FP+TN}\right)}$.

However, these metrics are only well-established for binary settings. Typically, Mean Absolute Error (MAE) is used for ordinal classification,

$$\mathrm{MAE} = \frac{1}{N}\sum_{i} |k_i - \hat{k}_i|.$$

where k_i and $\hat{k}_i$ are the real and predicted class for the i-th observation, respectively.

This metric suffers from two problems, however. It treats an ordinal variable as a cardinal variable, and the metric is biased when class imbalanced is present. The first problem is domain-dependent, but the second problem may be addressed using such metrics as Average Mean Absolute Error (AMAE), and Maximum Absolute Error (MMAE),

$$\mathrm{AMAE} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{MAE}_k, \tag{2}$$

$$\mathrm{MMAE} = \max\{\mathrm{MAE}_k \mid k = 1, \ldots, K\}, \tag{3}$$

where MAE_k means MAE is computed only for class k, i.e. $\mathrm{MAE}_k = \frac{1}{N}\sum_{i(k)} |k - \hat{k}_i|$, so that indexing traverses only class k. These metrics have also been used in Pérez-Ortiz et al. [12].

5.4 Models

The state of the art models hereby tested are SVOR [2] and RankSVM [6]. These models are then compared with F&H using SVM and binary RankSVM, respectively. Regularization is applied between these two family of models with $\lambda = 1$ being strong regularization and $\lambda = 0$ being no regularization.

Models are trained using gradient descent with the learning rate being $\eta = \frac{1}{\lambda i}$, with i being the number of iterations so far, as in Pegasos [13]. Data is normalized first using standard score.

5.5 Results for synthetic data

Initial results for the synthetic data may be found at Table 2. Parameters in bold represents the default values of the parameters. In each parameter group, the parameter is changed while all other parameters are fixed using the values specified in bold. The size of the training and evaluation data is the same, $N = 400$ using the geometric imbalance described before. Values in bold are the best scores, i.e. with the lowest AMAE score. The scores represent averages of five runs, with the bold values representing the minimum scores.

The parameters are angle variation between classes (Δa), range of noise in the class direction (ε), and range of noise in the orthogonal direction (ε'). See 5.1 for more details.

Ordinal RankSVM failed to overcome SVOR's performance, yet the proposed ensemble fares much better than its counterpart. As expected, when $a_k = 0\ \forall k$, the best results are when λ is high, forcing decision hyperplanes to be nearly parallel (i.e. equal to the base model). As Δa is increased,

Table 2 Results for various parameters of the synthetic data, using AMAE as the metric (lower is better)

		Regularization					Regularization				
		SVOR	F&H w/ SVM				RankSVM	F&H w/ RankSVM			
			$\lambda = 1$	$\lambda = 0.1$	$\lambda = 0.01$	$\lambda = 0$		$\lambda = 1$	$\lambda = 0.1$	$\lambda = 0.01$	$\lambda = 0$
Δa	0	0.087	0.091	0.091	0.083	0.100	0.170	**0.071**	0.075	0.074	0.072
	15	0.105	0.107	0.103	0.100	0.119	0.140	0.099	0.097	**0.097**	0.106
	30	0.127	0.123	0.121	0.099	0.116	0.181	0.121	0.109	0.099	**0.091**
	45	0.195	0.186	0.159	0.103	0.123	0.223	0.159	0.143	0.108	**0.096**
ε	0	0.018	0.010	0.002	0.001	0.022	0.046	0.002	0.001	**0.000**	**0.000**
	0.2	0.132	0.127	0.117	0.093	0.118	0.169	0.102	0.099	**0.085**	0.087
	0.4	0.350	0.350	0.339	0.344	0.356	0.365	0.336	0.336	**0.327**	**0.327**
ε'	0	0.090	0.087	0.076	0.076	0.101	0.100	0.072	0.071	**0.069**	0.070
	0.2	0.133	0.122	0.117	0.096	0.114	0.175	0.111	0.105	0.089	**0.078**
	0.4	0.191	0.182	0.155	0.107	0.119	0.268	0.165	0.130	0.108	**0.098**

then the smaller values of λ (which allows more flexible decision boundaries) result in improved performance.

The noise introduced worsen the performance of the models as expected, with the best results falling to the proposed ranking F&H approach. Surprisingly, even the the orthogonal noise (ϵ') tricks the models.

Further experiments were performed to evaluate the impact of the quality of the base model on the performance of the F&H ensemble. This was performed by (i) increasing ϵ only for the base model, and also by (ii) forcing the base model to learn from smaller samples of the training data. In both cases, the F&H ensemble seems resilient. As noise is increased, the error becomes more volatile, but the impact on the average AMAE across simulations is barely noticeable. This is explained by a shift in the λ that produces the minimum AMAE. As expected, the best λ becomes zero as the base model becomes more unreliable.

5.6 Results for real datasets

Results obtained for real datasets are presented in Tables 3 and 4, and evaluated using the two metrics described above: AMAE and MMAE. These results are averages for the 30 folds of stratified cross-validation. The best scores are

Table 3 Results using AMAE as the metric (lower is better)

	Regularization					Regularization				
	SVOR	F&H w/ SVM				RankSVM	F&H w/ RankSVM			
		$\lambda = 1$	$\lambda = 0.1$	$\lambda = 0.01$	$\lambda = 0$		$\lambda = 1$	$\lambda = 0.1$	$\lambda = 0.01$	$\lambda = 0$
toy	1.20	1.20	1.20	1.18	1.18	1.35	1.10	0.98	0.93	**0.91**
wisconsin5	**1.16**	1.21	1.21	1.24	1.24	1.23	1.22	1.25	1.30	1.26
wisconsin10	**2.44**	2.51	2.51	2.56	2.51	2.59	2.67	2.71	2.75	2.70
diabetes10	1.46	1.48	1.48	1.57	1.66	1.56	**1.29**	**1.29**	1.38	1.39
contact-lenses	**0.43**	**0.44**	**0.44**	**0.44**	**0.40**	**0.40**	0.49	0.48	0.48	**0.45**
ERA	1.49	1.41	1.41	1.51	1.66	1.49	**1.27**	**1.26**	**1.26**	**1.26**
diabetes5	0.85	0.84	0.84	0.87	0.91	0.69	**0.63**	**0.60**	**0.60**	**0.60**
auto5	0.43	0.41	0.41	0.42	0.42	0.42	0.39	0.36	**0.34**	**0.34**
LEV	0.73	0.65	0.65	0.66	0.66	0.65	0.59	0.59	**0.58**	**0.58**
auto10	1.01	0.78	0.78	0.89	0.95	0.80	0.74	0.72	**0.67**	**0.68**
SWD	0.61	0.61	0.61	0.61	0.62	0.61	**0.53**	**0.53**	0.54	0.54
machine10	1.08	1.01	1.01	1.04	1.43	1.05	**0.96**	**0.94**	**0.93**	**0.93**
machine5	0.52	**0.45**	**0.45**	0.46	0.60	0.49	**0.43**	**0.41**	**0.41**	**0.42**
ESL12vs3vs456...	0.61	**0.29**	**0.29**	0.49	0.62	0.34	0.32	0.31	0.32	0.33
ERA1vs23456vs...	1.06	0.88	0.88	1.14	1.34	0.66	**0.62**	**0.61**	**0.62**	0.63
Average	1.00	0.94	0.94	1.00	1.08	0.96	0.88	0.87	0.87	0.87
Winner	20%	20%	20%	7%	7%	7%	47%	47%	53%	60%

Table 4 Results using MMAE as the metric (lower is better)

	Regularization					Regularization				
	SVOR	F&H w/ SVM				RankSVM	F&H w/ RankSVM			
		$\lambda = 1$	$\lambda = 0.1$	$\lambda = 0.01$	$\lambda = 0$		$\lambda = 1$	$\lambda = 0.1$	$\lambda = 0.01$	$\lambda = 0$
toy	2.00	2.00	2.00	2.00	2.00	2.02	1.70	1.45	1.40	**1.38**
wisconsin5	1.88	1.90	1.90	1.91	1.98	**1.81**	**1.74**	**1.75**	1.84	**1.80**
wisconsin10	4.66	4.71	4.71	4.83	4.87	**4.31**	**4.47**	**4.54**	4.61	**4.49**
diabetes10	3.12	3.18	3.18	3.40	3.50	3.06	**2.58**	**2.66**	2.86	2.80
contact-lenses	**0.88**	**1.01**	**1.01**	**0.95**	**0.89**	**0.88**	**1.00**	**1.00**	1.06	**1.01**
ERA	2.47	2.26	2.26	2.52	2.99	2.25	**1.84**	**1.85**	**1.83**	**1.84**
diabetes5	1.48	1.50	1.50	1.48	1.52	1.24	**1.04**	**1.03**	**1.04**	**1.03**
auto5	1.00	1.00	1.00	1.00	1.00	0.83	0.83	**0.75**	**0.73**	**0.71**
LEV	1.37	1.30	1.30	1.29	1.28	**1.01**	**1.03**	**1.06**	**1.04**	**1.05**
auto10	2.97	2.05	2.05	2.48	2.67	**1.71**	**1.74**	**1.74**	**1.69**	**1.70**
SWD	1.11	1.12	1.12	1.10	1.10	1.00	**0.71**	0.73	0.77	0.78
machine10	2.97	3.02	3.02	**2.69**	3.62	3.17	2.98	**2.91**	**2.77**	**2.72**
machine5	1.12	**1.05**	**1.05**	1.06	1.35	1.12	**0.97**	**0.96**	**0.97**	**1.00**
ESL12vs3vs456...	1.07	0.68	0.68	0.98	1.09	0.72	**0.58**	**0.57**	**0.57**	**0.59**
ERA1vs23456vs...	2.10	1.51	1.51	2.16	2.77	**1.10**	**1.07**	**1.07**	**1.11**	1.11
Average	2.01	1.89	1.89	1.99	2.17	1.75	1.62	1.60	1.62	1.60
Winner	7%	13%	13%	13%	7%	40%	80%	87%	60%	80%

presented in bold, as are all scores which are statistically identical to the best based on a one-tailed paired different Student's *t*-test with a 95% confidence level hypothesis test. Notice therefore, more than one bold item is presented in each row, and the number of winners (last table row) does not sum to 100%.

The improvement in the last column is notorious relative to (a) the state-of-the-art SVOR and SVM family of models, and (b) to RankSVM as proposed in the previous paper, which by itself is already an improvement over the state-of-the-art methods. The results confirm the state-of-the-art in that ordinal RankSVM outperforms SVOR [6]. Notice the tables are ordered by ascending Imbalance Ratio (see Table 1). This is further corroborated by the correlations from Table 5, discussed in the next paragraph. Considering the datasets average of AMAE, SVOR has the worst results with 1.00 followed by ordinal RankSVM with 0.96 and then F&H with binary RankSVM with 0.87.

Table 5 performs Kendall τ rank correlations between the AMAE scores and the data characteristics from Table 1. The table shows AMAE error reducing as dataset size increases, which makes sense given the models have more data from which the underlying distribution can be inferred. However, it is interesting to note that this reduction seems to be inverse to the λ regularization being used, i.e. forcing parallel decision hyperplanes seems to work best for data with more observations.

This puzzling fact is possibly answered by the next row, showing that parallel hyperplanes work best for imbalance data, possibly due to the difficulty of estimating parameters associated to the minority classes, therefore, regularization transfers knowledge about the hyperplane slope from the other majority classes. There is a an even higher correlation between the overlap ratio (OR) and the AMAE scores, as suggested by [7].

Table 5 Correlation between dataset characterists and AMAE scores

	Regularization					Regularization				
	SVOR	F&H w/ SVM				RankSVM	F&H w/ RankSVM			
		$\lambda = 1$	$\lambda = 0.1$	$\lambda = 0.01$	$\lambda = 0$		$\lambda = 1$	$\lambda = 0.1$	$\lambda = 0.01$	$\lambda = 0$
N	− 0.01	− 0.01	− 0.01	− 0.07	− 0.03	0.03	− 0.01	− 0.03	− 0.03	− 0.01
IR	− 0.14	− 0.16	− 0.16	− 0.22	− 0.18	− 0.10	− 0.12	− 0.14	− 0.10	− 0.12
OR	0.26	0.20	0.20	0.26	0.30	0.14	0.16	0.18	0.14	0.16

6 Conclusion

Previous work by the authors has shown pairwise scoring ranking to be a benefit for class imbalance in the binary context [4], and then in the ordinal context [5, 6].

However, these later models suffered from class imbalance themselves, due to an internal imbalance of the pairwise comparisons being performed. This is addressed in this paper which allies Frank and Hall with a binary implementation of the model.

The results are able to reproduce the findings from the previous publication using the ordinal RankSVM classifier [6], and are improved further by the current proposal of using our binary RankSVM from [4] together with the Frank and Hall ensemble [8]. On average, the improvements are on the order of 12% relative to the previous publication, and 15% relative to a model from the state-of-the-art.

As future work, we would like to extend our approach by joint optimization of the hyperplanes and new ideas for fusing the outputs of the members of the ensemble, e.g. without optimizing thresholds but rather taking into account soft predictions. Moreover, a weighting-scheme could be introduced into the ranking approach, to pose more importance on pairs that belong to distant classes in the ordinal scale.

Acknowledgements This work was funded by the Project "NanoSTIMA: Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics/NORTE-01-0145-FEDER-000016" financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF), and also by Fundação para a Ciência e a Tecnologia (FCT) within PhD grant numbers SFRH/BD/122248/2016 and SFRH/BD/93012/2013.

References

1. Cardoso JS, Costa JF (2007) Learning to classify ordinal data: the data replication method. J Mach Learn Res 8((Jul)):1393–1429
2. Chu W, Keerthi SS (2007) Support vector ordinal regression. Neural Comput 19(3):792–815
3. Costa JFP, Sousa R, Cardoso JS (2010) An all-at-once unimodal svm approach for ordinal classification. In: 2010 Ninth international conference on machine learning and applications (ICMLA), IEEE, pp 59–64
4. Cruz R, Fernandes K, Cardoso JS, Costa JFP (2016) Tackling class imbalance with ranking. In: 2016 International joint conference on neural networks (IJCNN), IEEE, pp 2182–2187
5. Cruz R, Fernandes K, Costa JFP, Ortiz MP, Cardoso JS (2017) Combining ranking with traditional methods for ordinal class imbalance. In: International work-conference on artificial neural networks. Springer, Cham, pp 538–548
6. Cruz R, Fernandes K, Costa JFP, Ortiz MP, Cardoso JS (2017) Ordinal class imbalance with ranking. In: Iberian conference on pattern recognition and image analysis. Springer, Cham, pp 3–12
7. Denil M, Trappenberg TP (2010) Overlap versus imbalance. In: Canadian conference on AI. Springer, pp 220–231
8. Frank E, Hall M (2001) A simple approach to ordinal classification. In: Machine learning: ECML 2001, pp 145–156
9. Gutiérrez PA, Pérez-Ortiz M, Sanchez-Monedero J, Fernández-Navarro F, Hervas-Martinez C (2016) Ordinal regression methods: survey and experimental study. IEEE Trans Knowl Data Eng 28(1):127–146
10. Herbrich R, Graepel T, Obermayer K (1999) Support vector learning for ordinal regression. In: Ninth international conference on artificial neural networks ICANN 99, vol 1, Edinburgh, pp 97–102
11. Li L, Lin HT (2007) Ordinal regression by extended binary classification. In: Advances in neural information processing systems, pp 865–872
12. Pérez-Ortiz M, Gutiérrez PA, Hervás-Martínez C, Yao X (2015) Graph-based approaches for over-sampling in the context of ordinal regression. IEEE Trans Knowl Data Eng 27(5):1233–1245
13. Shalev-Shwartz S, Singer Y, Srebro N, Cotter A (2011) Pegasos: primal estimated sub-gradient solver for svm. Math Program 127(1):3–30
14. Wang H, Shi Y, Niu L, Tian Y (2017) Nonparallel support vector ordinal regression. IEEE Trans Cybern 47(10):3306–3317