Fine-to-Coarse Ranking in Ordinal and Imbalanced Domains: An Application to Liver Transplantation

María Pérez-Ortiz[1(✉)], Kelwin Fernandes[2,3], Ricardo Cruz[3], Jaime S. Cardoso[2,3], Javier Briceño[4], and César Hervás-Martínez[5]

[1] Department of Quantitative Methods, Universidad Loyola Andalucía, Córdoba, Spain
mariaperez@uloyola.es
[2] Faculty of Engineering, University of Porto, Porto, Portugal
kafc@inesctec.pt, jaime.cardoso@inesctec.pt

[3] INESC TEC, Porto, Portugal
rpcruz@inesctec.pt
[4] Liver Transplantation Unit, Reina Sofia Hospital, Córdoba, Spain
[5] Department of Computer Science and Numerical Analysis, University of Córdoba, Córdoba, Spain
chervas@uco.es

Abstract. Nowadays imbalanced learning represents one of the most vividly discussed challenges in machine learning. In these scenarios, one or some of the classes in the problem have a significantly lower a priori probability, usually leading to trivial or non-desirable classifiers. Because of this, imbalanced learning has been researched to a great extent by means of different approaches. Recently, the focus has switched from binary classification to other paradigms where imbalanced data also arise, such as ordinal classification. This paper tests the application of learning pairwise ranking with multiple granularity levels in an ordinal and imbalanced classification problem where the aim is to construct an accurate model for donor-recipient allocation in liver transplantation. Our experiments show that approaching the problem as ranking solves the imbalance issue and leads to a competitive performance.

Keywords: Imbalanced data · Ranking · Ordinal classification · Oversampling

This work has been subsidized by the TIN2014-54583-C2-1-R and the TIN2015-70308-REDT projects of the Spanish Ministerial Commission of Science and Technology (MINECO, Spain), FEDER funds (EU), the PI-0312-2014 project of the "Fundación pública andaluza progreso y salud" (Spain), the PI15/01570 project ("Proyectos de Investigación en Salud"), and also by NanoSTIMA: Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics/NORTE-01-0145-FEDER-000016.

I. Rojas et al. (Eds.): IWANN 2017, Part II, LNCS 10306, pp. 525–537, 2017.
DOI: 10.1007/978-3-319-59147-6_45

1 Introduction

Liver transplantation, although representing nowadays a widely-accepted and successful treatment for patients who present terminal liver disease, is hampered by the low availability of suitable donors. Several strategies have been considered over the years to construct a system to prioritize recipients on the waiting list and optimise the utility of the organ, but most of them only consider characteristics of either donors or recipients, therefore obviating the potential compatibility between these. This paper proposes a novel donor-recipient liver allocation system based on machine learning by means of a dataset comprised of donor-recipients pairs from different centres (seven Spanish transplantation units and the London King's College hospital).

The problem tackled in this paper is imbalanced, as it is common in biomedicine and real-world applications, meaning this that there exist one or several classes that are under-represented in the dataset, which usually leads the classification strategy to a non-appropriate trivial final model [4,22]. In this case, the dependent variable is the time leading up to graft loss, where patients are monitored for a year after transplantation. The classes of the problem are: less than 15 days, between 15 and 90 days, between 90 and 365 days and more than a year (showing then an ordinal nature). The number of transplants that result in graft loss are in this case, a significant minority with respect to successful transplants, however, these cases are the most interesting ones for constructing a proper model for organ allocation.

To deal with class imbalance, different approaches have been proposed over the years, mainly data-based approaches [4,12] (such as over-sampling the minority class or under-sampling the majority one) or algorithmic approaches [3] (e.g. cost-sensitive learning or post-processing strategies). Since imbalanced data does not only arise in binary standard classification domains, there are different extensions of these initially proposed algorithms to other learning paradigms being researched [19,23].

In the case of ordinal classification, the different categories of the class to predict follow a given ranking order, but not a cardinal order. This must be considered in the different stages of the learning process; such as training the algorithm, measuring the performance and applying preprocessing techniques (e.g. over-sampling or under-sampling).

As said, from all the techniques that have been proposed to deal with class imbalance, over-sampling and under-sampling approaches can be highlighted, because they have been shown to improve classifier performance over imbalanced datasets and do not depend on a specific classification approach. In this sense, previous studies suggest that over-sampling could be more powerful than under-sampling [16], specially for highly imbalanced and small datasets. This could be due to the potential loss of useful information when performing under-sampling. The main problem, however, with over-sampling approaches is that most of these methods assume a distribution of minority samples which may not hold in reality, resulting in synthetic patterns lying in inadequate areas of the dataset [20]. Concerning algorithmic approaches, cost-sensitive learning has been shown to result in over-fitting [10,20] and although hybrid and ensemble algorithms are common and usually successful, they rely on a specific method and are difficult to optimise. Recently, a novel proposal has joined the set of strategies available to deal with class imbalance, the use of rank learners, which has shown very promising results for this matter [6,7]. This paper explores the use of this type of algorithms in the context of ordinal and imbalanced data and compares it to other proposals in the literature.

In our case, the use of rankers is natural, given that: (1) the ordinal classes to predict (i.e. that represent the time leading up to graft failure) come from a continuous latent variable which represents the exact number of days[1] and (2) from the definition of the problem tackled our aim is to construct an accurate organ allocation system, that ranks patients according to their suitability for the transplant. This information can be used to construct the model, introducing valuable information about the rank of the patterns, which is lost when using coarse labels. The experiments performed in this paper compare different rank-based learners with several ordinal classifiers and techniques to deal with class imbalance, showing that: (1) the problems that present such highly imbalanced nature require the use of specially designed techniques to avoid trivial classifiers; (2) rank-based learners represent a suitable option to deal with imbalanced data; (3) these strategies show comparable results to the use of other approaches to deal with class imbalance, and (4) the use of fine labels (as opposed to coarse labels) can satisfactorily complement the classification leading to more robust results.

The paper is organised as follows: Sect. 2 shows a description of the dataset and methodology used; Sect. 3 describes the experimental study and analyses the results obtained; and finally, Sect. 4 outlines some conclusions.

2 Materials and Methodology

2.1 Dataset Description

A multi-centred retrospective analysis was made of 7 Spanish Liver transplant units. Recipient, donor, retrieval and transplant characteristics were reported at the time of transplant. Patients undergoing partial, split or living-donor liver transplantation and patients undergoing combined or multi-visceral transplants were excluded from the study. Liver transplantation units were homogeneously distributed throughout Spain. The Spanish dataset constructed has 634 patterns (donor-recipient pairs) corresponding to the years 2007 and 2008. The proportion of combined transplant was 2.3% in both cohorts. The proportion of partial grafts was 0.9% and 9.1% in the Spanish and British cohort, respectively. The few cases of combined transplantation were those of liver and kidney, which, in several series, have been reported to not decrease the outcome of the liver graft.

[1] Note, however, that a simple regression analysis is not feasible because of the high number of organs which survived the 365 day threshold (for which, we do not have more information).

In addition, the dataset was completed with information about donor-recipient pairs from the King's College Hospital (London), to perform a supranational study of donor-recipient allocation in liver transplantation. To obtain a similar number of patterns, only reported pairs of recipients over eighteen years of age between January 2002 and December 2010 were included. A dataset containing 858 English donor-recipient pairs was collected. In order to merge the datasets, several variables were selected, 16 recipient variables, 17 donor variables and 5 surgically related variables, as can be seen in [18]. All patients were followed from the date of transplant until either death, graft loss or completion of the first year after the liver transplant. The final dataset was comprised of 1406 patterns.

To solve the donor-recipient matching problem, the dependent variable is the class label which is equal to 1 when representing graft loss up to the first 15 days after the transplant, equal to 2 if the loss occurs between 15 days and 3 months, equal to 3 when the loss is after 3 months and before a year, and, finally, the last class corresponds to the patterns which do not present graft loss after the first year and is represented by label 4. The variables selected for the dataset can be seen in [18].

The choice of class limits for the dataset were not arbitrary (15 days, 3 months and a year); in addition to being considered as the most pertinent, [18] shows that the cumulative frequency slope of the graft loss curve changes strongly somewhere around those class limits. An important point is the limit located at 15 days since it is defined by experts as a critical point for survival or loss. In this case, the application of a regression-based technique is not suitable for the problem, due to the high number of points belonging to the more than 1 year category, which do not incorporate any knowledge about the real value of the number of days until either graft loss or death. The class distribution of the dataset is {76,76,62,1223}, which shows a highly imbalanced nature.

2.2 Methodology

The ordinal regression problem consists of predicting the label y for an input vector $\mathbf{x}$, where $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ and $y \in \mathcal{Y} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_Q\}$, i.e. $\mathbf{x}$ is in a K-dimensional input space and y is in a label space of Q different labels corresponding to the categories. The objective is to find a classification rule or function $f \colon \mathcal{X} \to \mathcal{Y}$ to predict the labels of new patterns, given a training set of N points, $D = \{(\mathbf{x}_i, y_i), i = 1, \ldots, N\}$. A natural label ordering is included for ordinal regression, $\mathcal{C}_1 \prec \mathcal{C}_2 \prec \ldots \prec \mathcal{C}_Q$, where $\prec$ is an order relation given by the nature of the classification problem. Many ordinal regression measures and algorithms consider the rank of the label, i.e. the position of the label in the ordinal scale, which can be expressed by the function $\mathcal{O}(\cdot)$, in such a way that $\mathcal{O}(\mathcal{C}_q) = q$, $q = 1, \ldots, Q$. The assumption of an order between class labels makes that two different elements of $\mathcal{Y}$ could always be compared by using the relation $\prec$, which is not possible under the nominal classification setting.

Rank Learners for Ordinal Classification. The *multipartite ranking* problem is a generalisation of the well-known bipartite ranking one. ROC analysis, which evaluates the ability of classifiers to sort positive and negative instances in terms of the area under the ROC curve, is a clear example of training a binary classifier to perform well in a bipartite ranking problem. Multipartite ranking can be seen as an intermediate point between ranking and sorting. It is similar to ranking because training patterns are labelled with one of Q ordered ratings ($Q = 2$ for bipartite ranking), but here the goal is to learn from them a ranking function able to induce a total order in accordance with the given training ratings [9,21], which is similar to sorting. The objective of multipartite ranking is to obtain a classifier which ranks "high" classes ahead of "low" classes (in the ordinal scale), being this a refinement of the order information provided by an ordinal classifier, as the latter does not distinguish between objects within the same category. The relationship between multipartite ranking and ordinal classification is discussed in [9]. An ordinal regression classifier can be used as a ranking function by interpreting the class labels as scores [13]. However, this type of scoring will produce a large number of ties (which is not desirable for multipartite ranking). On the other hand, a multipartite ranking function $f(\cdot)$ can be turned into an ordinal classifier by deriving thresholds to define an interval for each class, but how to find the optimal thresholds is an open issue.

We propose to consider pairwise scoring rankers for the class imbalance problem. Ranking algorithms have been found to be very competitive classifiers for classification [6]. And the crucial property is that there is no class imbalance when performing pairwise ranking, since each observation of a class is compared to every observation of the other class, being the associated learning process balanced.

Scoring pairwise rankers are a type of rankers, where each observation $\mathbf{x}_i$ is compared against all others $\mathbf{x}_j$, and if $\mathbf{x}_i \succ \mathbf{x}_j$, then the model learns a scoring function s so that if $\mathbf{x}_i \succ \mathbf{x}_j$ then $s(\mathbf{x}_i) > s(\mathbf{x}_j)$, with $s\colon X \rightarrow \mathbb{R}$.

This scoring function then needs to be converted back to ordinal classes. A threshold strategy is proposed. Let s_i be the ordered score of observation i and q_i be the true class, we search the threshold left-to-right by invoking the function min_error with initial parameters $(s_0, q_0, 0)$, where:

$$\text{min\_error}(s_i, q_i, \hat{q}) = \begin{cases} \varepsilon_{q_i\hat{q}}, & \text{when } i = N, \\ \min\left\{\varepsilon_{q_i\hat{q}} + \text{min\_error}(s_{i+1}, q_{i+1}, \hat{q}),\ \text{min\_error}(s_i, q_i, \hat{q}_{i+1})\right\} \end{cases}$$

Several cost matrices ε can be used. For instance:

- **Homogeneous:** $\varepsilon_{q\hat{q}} = \{1 \text{ if } q \neq \hat{q} \text{ and } 0 \text{ otherwise} \mid \forall q, \hat{q}\}$
- **Absolute costs:** $\varepsilon_{q\hat{q}} = \{|q - \hat{q}| \mid \forall q, \hat{q}\}$
- **Inverse class frequency:** $\varepsilon_{q\hat{q}} = \left\{\frac{N}{QN_q+1} \text{ if } q \neq \hat{q} \text{ and } 0 \text{ otherwise} \mid \forall q, \hat{q}\right\}$, where N_q is the number of patterns in class $\mathcal{C}_q$.

So far the underlying model has not been discussed; several adaptations have been presented in the literature: from neural networks [2] to gradient boosting [24]. In this work, the underlying model used is an SVM trained in the space of differences, an adaptation of [14]. The space of differences is constructed by

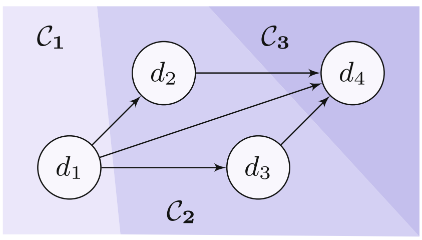
$\mathbf{x}_{mn}^{(ab)} = \mathbf{x}_m^{(a)} - \mathbf{x}_n^{(b)}$ with $\mathcal{C}_a < \mathcal{C}_b$, for every two observations indexed by m and n. The pairs are imbalance however, so each pair is re-weighted by a factor of $\prod_{q=1}^{Q} N_q/(N_a N_b)$, which corresponds to the inverse class frequency.

Fine-to-Coarse Ranking. Our learning problem is framed as an ordinal regression problem with two levels of granularity on the target variable $\mathcal{Y}_{\text{fine}} = \{\mathcal{C}_1^{\text{fine}}, \mathcal{C}_2^{\text{fine}}, \ldots, \mathcal{C}_{Qf}^{\text{fine}}\}$ and $\mathcal{Y}_{\text{coarse}} = \{\mathcal{C}_1^{\text{coarse}}, \mathcal{C}_2^{\text{coarse}}, \ldots, \mathcal{C}_{Qc}^{\text{coarse}}\}$, such that $\mathcal{Y}_{\text{fine}}$ can be monotonously partitioned in Qc disjoint groups.

This is a very common setting and can be seen in a wide range of scenarios. For instance, stundents receive a continuous fine-grained mark (e.g. 0–100%) which is usually under-segmented into a smaller number of intervals (e.g. fail, pass, pass with honors). Another example can be observed in ranking companies being broadly categorized as small-medium-large enterprises based on a fine-grained scale given by the number of employees, annual turnover, etc.

In this sense, coarse labels are a semantic abstraction of the original phenomenon being quantified. Despite this abstraction may simplify the analysis of each group, it may impose a significant loss of information when building rankers. For instance, for the specific problem tackled in this work, the number of comparisons used for training in a coarse scheme is restricted to patients on different intervals, while on a fine scheme we can distinguish patients that survived n days from patients that survived $n + 1$ days (see Fig. 1).

Thus, we propose to train the ranking model with the entire original information (fine labels) in order to obtain a sound and stable ranker. On the other hand, the transformation from ranking to ordinal classification is done using the ordinal classes (coarse labels).



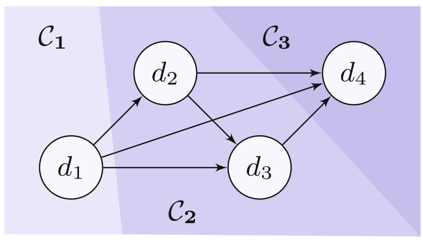


Fig. 1. Comparison of Coarse Ranking and Fine-to-Coarse Ranking. Transitive edges are not shown to improve readability.

Over-Sampling in Ordinal Domains. As stated before, the ordinal nature of the classes in ordinal classification should be taken into account for all the stages of the learning machine (e.g. for preprocessing the data). To compare the impact of rank-based learners, this paper also considers two over-sampling techniques in the context of ordinal classification and regression [19,23]. The main intuition behind these methods is that the ordering structure of the classes can be exploited when generating new synthetic patterns. A synthetic pattern in this case corresponds to a virtual donor-recipient pair, which we create to balance the class distribution and make the classifier pay more attention to minority classes. These virtual pairs are created using the information of other pairs, so new synthetic patterns are not totally virtual, but rather based on the combination of two donors and two recipients.

The strategies tested in this paper are:

- Over-sampling for regression problems (OR): The idea of using over-sampling in the presence of real-valued outputs presented in [23] is used in this paper, as an alternative when fine labels are given (instead of coarse categories). In this case, new synthetic patterns are created using a convex combination of two neighbours, where the new label is created also by convex combination of the label of the two patterns.
- Ordinal graph-based over-sampling (OGO): The ordinal nature of the data is exploited considering a neighborhood graph between the classes, which aims to capture the underlying manifold of the ordinal labelling space. New patterns are generated on the paths that preserve the ordinal structure of this manifold and create a spatial continuity on the input space.

Further details about these methods can be consulted in [19,23].

3 Experiments

3.1 Methodologies Tested

The experiments in this paper have been designed to compare several methodologies with our proposal of using rank-based learners as an alternative to imbalanced classification:

- Support vector machine using the one-versus-one approach (SVM) [15], a state-of-the-art nominal classification methodology. Both linear (LSVM) and non-linear (SVM) versions are employed.
- Cost-sensitive SVM [3], which poses a higher cost to minority classes (using the imbalance ratio of the class).
- Support Vector Machine for Ordinal Regression with Implicit Constraints (SVORIM) [5], a state-of-the-art SVM-based technique for ordinal classification.
- The proportional odds model (POM) [17], i.e. a standard ordinal logistic regression method which is widely used in the literature.
- Rank, the strategy presented in Sect. 2.2, where the ordinal classification problem is transformed to a learning to rank problem.
- F2C-Rank (also presented in Sect. 2.2) where the days leading up to graft failure are used to construct the ranking between the patterns.

As said, different threshold optimisation techniques are also tested for both the Rank and F2C-Rank strategies, in order to transform the ranking to discrete labels, as well as two over-sampling approaches (OGO and OR, depending on the nature of the label used).

3.2 Evaluation Metrics

Several measures can be considered for evaluating ordinal classifiers. The most common ones in machine learning are the Mean absolute error (MAE) and the accuracy (Acc) [11]. However, these measures may not be the best option, for example, when measuring performance in the presence of class imbalance [1], and/or when the costs of different errors vary markedly. The accuracy (Acc) is defined by:

$$Acc = \frac{100}{N} \sum_{i=1}^{N} (I(y_i^* = y_i)),$$

where $I(\cdot)$ is the zero-one loss function, y_i is the desired output for pattern $\mathbf{x}_i$, y_i^* is the prediction of the model and N is the total number of patterns in the dataset. However, this metric does not take the order of the categories into account, and it is not recommended for imbalanced datasets.

The average mean absolute error ($AMAE$) [1] is the mean of MAE classification errors throughout the classes, where MAE is the average absolute deviation of the predicted class from the true class (in number of categories on the ordinal scale). It is able to mitigate the effect of imbalanced class distributions. Let MAE_j be the MAE for a given j-th class:

$$MAE_j = \frac{1}{N_j} \sum_{i=1}^{N_j} |\mathcal{O}(y_i) - \mathcal{O}(y_i^*)|, \; 1 \leq j \leq Q,$$

where $\mathcal{O}(\mathcal{C}_j) = j$, $1 \leq j \leq Q$, i.e. $\mathcal{O}(y_j)$ is the order of class label y_j. Then, the $AMAE$ measure can be defined in the following way:

$$AMAE = \frac{1}{Q} \sum_{j=1}^{Q} MAE_j.$$

The Maximum Mean Absolute Error ($MMAE$), which corresponds to the MAE value considering only the patterns from the class with the greatest difference between true values as compared to the predicted ones:

$$MMAE = \max \{MAE_j; j = 1, \ldots, Q\},$$

where MAE_j is the MAE value considering only the patterns from the j-th class. This measure was recently proposed [8] and it is very interesting, since a low $MMAE$ represents a low error for all classes. MAE values are between 0 and $Q - 1$, and so are $AMAE$ and $MMAE$.

These four performance metrics briefly summarise the most important aspects of confusion matrices when dealing with ordinal and imbalanced data: Acc giving an idea of the global performance, MAE representing overall ordinal errors and $AMAE$ and $MMAE$ reflecting the magnitude of errors in the ordinal scale (the former the mean deviation and the latter the deviation associated to the worst classified class).

3.3 Experimental Setting

For evaluating the results, a stratified 10-fold technique has been used to divide the data, and the results have been taken as the mean and standard deviation.

The parameters for all methods have been chosen using a nested 5-fold cross-validation over the training set (independently of the 10-fold technique). The final parameter combination was the one which obtained, in mean, the best average performance for the 5 validation sets of this nested 5-fold cross-validation, where the metric used was the $AMAE$. The test sets were never used during model selection. The kernel selected for all the non-linear kernel methods was the Gaussian one, $K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma^2}\right)$, where σ is the kernel width. For every tested kernel method, the kernel width was tuned within the range $\sigma \in \{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$, as well as the cost parameter associated to SVM-based methods, $C \in \{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$.

3.4 Results

Table 1 shows the results obtained for the dataset considered with the methods previously mentioned. Several conclusions can be extracted from this table: Firstly, the complexity of the problem considered can be appreciated, as most methods obtain trivial models (accurate but that almost always predict the survival class, as indicated by the values of AMAE and MMAE of 1.5 and 3 respectively). Secondly, the choice of a linear or nonlinear model does not seem to have an important impact on the performance (as can be seen comparing the results of LSVM and SVM or CS-LSVM and CS-SVM). Thirdly, cost-sensitive approaches do not present an acceptable performance, confirming this that over-sampling should be preferred. In this regard, over-sampling helps the method to focus on the classification of the minority class and avoid trivial classifiers. The sole use of an ordinal method is not enough to obtain an appropriate classifier, as can be seen when analysing the results of LSVORIM (very competitive results for imbalance-nature metrics but a extremely poor overall performance, even worse than a random classifier). Concerning rank-based learners, these are seen to obtain a competitive performance and reach the results of other state-of-the-art classifiers (such as POM when combined with an ordinal over-sampling strategy). The use of fine labels help the method to optimise the results even further, specially for Acc, MAE and $AMAE$, obtaining worse results for $MMAE$, but a promising trade-off in all metrics. The combination of a rank-based learner with over-sampling is not satisfactory, which could mean that the sole use of a rank

Table 1. Average and standard deviation results for the test sets

Method	Acc	MAE	$AMAE$	$MMAE$
Nominal and ordinal classifiers				
LSVM	**85.11 ± 0.34**	*0.304 ± 0.004*	1.500 ± 0.000	3.000 ± 0.000
CS-LSVM	85.04 ± 0.47	0.306 ± 0.008	1.501 ± 0.002	3.000 ± 0.000
SVM	**85.11 ± 0.34**	*0.304 ± 0.004*	1.500 ± 0.000	3.000 ± 0.000
CS-SVM	**85.11 ± 0.34**	**0.303 ± 0.005**	1.500 ± 0.000	3.000 ± 0.000
LSVM+OGO	**85.11 ± 0.79**	**0.303 ± 0.017**	1.500 ± 0.000	3.000 ± 0.000
LSVORIM	20.20 ± 31.69	1.435 ± 0.196	**1.102 ± 0.196**	**2.185 ± 0.393**
LSVORIM+OGO	54.82 ± 31.97	0.994 ± 0.822	1.420 ± 0.179	2.566 ± 0.402
POM	*84.97 ± 0.34*	0.304 ± 0.006	1.500 ± 0.001	3.000 ± 0.000
POM+OGO	63.33 ± 3.12	0.540 ± 0.049	1.410 ± 0.086	2.562 ± 0.179
Rank-based learners				
Rank	64.37 ± 19.29	0.804 ± 0.439	1.422 ± 0.081	*2.416 ± 0.541*
F2C-Rank	67.75 ± 12.85	0.669 ± 0.281	*1.407 ± 0.096*	2.452 ± 0.511
Rank+OGO	69.74 ± 20.14	0.600 ± 0.399	1.443 ± 0.109	2.624 ± 0.497
F2C-Rank+OR	84.69 ± 0.45	0.310 ± 0.012	1.493 ± 0.021	2.975 ± 0.079
Comparison between threshold optimisation strategies for ranking				
Rank-Inv	64.37 ± 19.29	0.804 ± 0.439	1.422 ± 0.081	2.416 ± 0.541
Rank-Unif	84.55 ± 1.05	0.314 ± 0.018	1.495 ± 0.022	3.000 ± 0.000
Rank-Abs	85.04 ± 0.36	0.306 ± 0.010	1.489 ± 0.026	3.000 ± 0.000
F2C-Rank-Inv	67.75 ± 12.85	0.669 ± 0.281	1.407 ± 0.096	2.452 ± 0.511
F2C-Rank-Unif	*84.97 ± 0.32*	0.306 ± 0.009	1.495 ± 0.020	3.000 ± 0.000
F2C-Rank-Abs	84.83 ± 0.26	0.310 ± 0.010	1.499 ± 0.011	3.000 ± 0.000

The best result is in **bold** face and the second best result is in *italics*.

learner is enough to deal with the imbalanced nature of the data. Specially, the use of OR (as opposed to F2C-Rank) deteriorates the results and leads almost to a trivial classifier, which could be due to the over-sampling strategy itself. Finally, considering the use of different threshold optimisation strategies, the use of the inverse function (which takes into account the imbalanced nature of the data) presents the best results (which are the ones used for comparison in the rest of the experiments).

Note that the developed computational models are used as a decision support system. With MELD (one of the current assignation methodologies used worldwide) donors are generally assigned to the candidates at greatest-risk, a policy that does not allow the transplant team to do the matching according to the principles of fairness and survival benefit [18]. The method proposed here for organ allocation seeks to minimise futile liver transplantation, giving primary attention to patients with the best predicted lifetime gained due to transplantation. Note that although rankers have been tested in this case using classification metrics, they have shown to produce a better ordering between the classes, meaning that the output of the ranking algorithm could be used for the construction of the allocation system.

4 Conclusions

A practical case of imbalance ordinal classification is analysed in this paper: a dataset consisting of liver transplant survival information from eight transplantation units. The survival information is given in the number of days up to either graft loss or death or one year (365). A simple regression would not be appropriate due to the non-linear relation between the independent and dependent variables, and the boundedness nature of the problem (0–365). Typically, survival information would be discretised in classes prior to the application of classification classifier.

The traditional approach is compared against ranking, in particular pairwise scoring ranking. This model family has been found to produce good results in an imbalance context [6]. The expressiveness of ranking avoids the prior discretisation of the survival variable, which has been named as fine labels. A scoring ranking model predicts a score for each observation, which indicates the ranking order against the others. This ranking score must then be converted back to the original discretised class-space, which we have called coarse labels.

The suggested approach is contrasted with traditional ordinal classification methods: vanilla SVM, SVORIM and POM, with both linear and RBF kernels. Since the dataset is imbalanced, and the classifier is evaluated using imbalance metrics, these classifiers are also evaluated by prior over-sampling and by introducing cost-sensitivity. The proposed method is also tested against ranking trained only using coarse labels.

In the experimental setting, the proposed ranking method is found to be highly competitive, especially when considering the MMAE and AMAE metrics, which are imbalance-sensitive metrics. It was found that over-sampling is always preferred to introducing inverse frequency costs in order to balance the classes. Interestingly, this was not verified for ranking. Ranking performs generally competitively against traditional approaches, even without over-sampling. The proposed ranking approach, which makes use of fine labels, presents a promising trade-off in all metrics resulting in a final model that could be useful as a decision support system for the medical community.

References

1. Baccianella, S., Esuli, A., Sebastiani, F.: Evaluation measures for ordinal regression. In: Proceedings of the Ninth International Conference on Intelligent Systems Design and Applications, ISDA 2009, pp. 283–287 (2009)
2. Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G.: Learning to rank using gradient descent. In: Proceedings of the 22nd international Conference on Machine Learning - ICML 2005, pp. 89–96. ACM, New York (2005)

3. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**(3), 27 (2011)
4. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
5. Chu, W., Keerthi, S.S.: Support vector ordinal regression. Neural Comput. **19**, 792–815 (2007)
6. Cruz, R., Fernandes, K., Cardoso, J.S., Costa, J.F.P.: Tackling class imbalance with ranking. In: 2016 International Joint Conference on Neural Networks (IJCNN), pp. 2182–2187, July 2016
7. Cruz, R., Fernandes, K., Pinto Costa, J.F., Perez Ortiz, M., Cardoso, J.S.: Ordinal class imbalance with ranking. In: Rojas, I., et al. (eds.) IWANN 2017, Part II. LNCS, vol. 10306, pp. 538–548. Springer, Cham (2017)
8. Cruz-Ramírez, M., Hervás-Martínez, C., Sánchez-Monedero, J., Gutiérrez, P.A.: Metrics to guide a multi-objective evolutionary algorithm for ordinal classification. Neurocomputing **135**, 21–31 (2014)
9. Fürnkranz, J., Hüllermeier, E., Vanderlooy, S.: Binary decomposition methods for multipartite ranking. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009. LNCS, vol. 5781, pp. 359–374. Springer, Heidelberg (2009). doi:10.1007/978-3-642-04180-8_41
10. Galar, M., Fernández, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **42**(4), 463–484 (2012)
11. Gutiérrez, P.A., Pérez-Ortiz, M., Sánchez-Monedero, J., Fernandez-Navarro, F., Hervás-Martínez, C.: Ordinal regression methods: survey and experimental study. IEEE Trans. Knowl. Data Eng. **28**(1), 127–146 (2016)
12. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. **9**(21), 1263–1284 (2009)
13. Herbrich, R., Graepel, T., Obermayer, K.: Support vector learning for ordinal regression. In: International Conference on Artificial Neural Networks, pp. 97–102 (1999)
14. Herbrich, R., Graepel, T., Obermayer, K.: Large margin rank boundaries for ordinal regression. Advances in Large Margin Classifiers, pp. 115–132. MIT Press, Cambridge (2000)
15. Hsu, C.W., Lin, C.J.: A comparison of methods for multi-class support vector machines. IEEE Trans. Neural Netw. **13**(2), 415–425 (2002)
16. Japkowicz, N., Stephen, S.: The class imbalance problem: a systematic study. Intell. Data Anal. **6**(5), 429–449 (2002)
17. McCullagh, P.: Regression models for ordinal data. J. Roy. Stat. Soc. **42**(2), 109–142 (1980)
18. Pérez-Ortiz, M., Cruz-Ramírez, M., Ayllón-Terán, M., Heaton, N., Ciria, R., Hervás-Martínez, C.: An organ allocation system for liver transplantation based on ordinal regression. Appl. Soft Comput. **14**(Part A), 88–98 (2014)
19. Pérez-Ortiz, M., Gutiérrez, P., Hervás-Martínez, C., Yao, X.: Graph-based approaches for over-sampling in the context of ordinal regression. IEEE Trans. Knowl. Data Eng. **27**(5), 1233–1245 (2015)
20. Pérez-Ortiz, M., Gutiérrez, P.A., Tino, P., Hervás-Martínez, C.: Oversampling the minority class in the feature space. IEEE Trans. Neural Netw. Learn. Syst. **27**(9), 1947–1961 (2016)

21. Rajaram, S., Agarwal, S.: Generalization bounds for k-partite ranking. In: Proceedings of the Seventeenth Annual Conference on Neural Information Processing Systems (NIPS2005), pp. 28–23 (2005)
22. Tang, Y., Zhang, Y.Q., Chawla, N.V., Krasser, S.: SVMs modeling for highly imbalanced classification. IEEE Trans. Syst. Man Cybern. B Cybern. **39**(1), 281–288 (2009)
23. Torgo, L., Ribeiro, R.P., Pfahringer, B., Branco, P.: SMOTE for regression. In: Correia, L., Reis, L.P., Cascalho, J. (eds.) EPIA 2013. LNCS (LNAI), vol. 8154, pp. 378–389. Springer, Heidelberg (2013). doi:10.1007/978-3-642-40669-0_33
24. Zheng, Z., Chen, K., Sun, G., Zha, H.: A regression framework for learning ranking functions using relative relevance judgments. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 287–294. ACM (2007)