Ordinal model reuse and selection for a varying number of categories

Adolfo Martínez-Usó, José Hernández-Orallo, Cèsar Ferri, and María José Ramírez-Quintana

DSIC, Universitat Politècnica de València
Camí de Vera s/n, 46022, València, Spain
{admarus,jorallo,cferri,mramirez}@dsic.upv.es

Abstract. Ordinal classification or ordinal regression is the supervised learning problem of predicting categories that have an ordered arrangement. Performance metrics are usually understood in terms of this ordinality and models are meant to make the fewer errors in these terms. In many scenarios, the number of bins of an ordinal regression problem vary from training to deployment. In the general case, we may have learnt a model with data for which the output is numeric (a traditional regression problem) and then we need to apply the model to a situation where ordinal categories are more meaningful. During training, the number of categories is not known, so we need to analyse models that perform well in a range of contexts, where the context is defined as the number of categories or bins. This is a common situation in sentiment analysis. In this work we analyse how a regression model is reframed (i.e., reused) for a range of contexts (number of bins) and also how we can use context plots to select the best models for particular contexts ranges, resulting in a hybrid model. A series of experiments using a real-world sentiment dataset and 25 regression datasets analyses if this region-wise hybrid model is advantageous over a single global model.

Keywords: Ordinal classification, ordinal regression, reframing, context change, binarisation problem

1 Introduction

In machine learning, predictive techniques have the aim of constructing patterns and models from past data to be applied to new collections of data. In this framework, the learning process is organised into two stages: the training stage, in that the model is built, and the deployment stage, in that the model is applied. More often, in real-world settings the operating context (that is, the set of characteristics and knowledge about where the model is learnt or applied) can differ from the training to deployment. For instance, in sentiment analysis, a coarse-grained number of categories for one context may be insufficient for other contexts, being generally preferred to have the possibility to manage more fine-grained sentiment distinctions [5].

In recent years, Ordinal Regression (OR) has witnessed an increased interest among the information retrieval community, mainly because of knowledge collected from product reviews, movie ratings, opinion mining or, in particular, sentiment analysis [12]. This knowledge is nowadays an important part of the marketing strategy of many companies [6]. Data collected for these tasks are often associated to a score of user/consumer satisfaction or of the sentiment reaction.

This paper focuses on OR, presenting a methodology that reuses the trained model when the number of bins can change depending on the application, i.e. the number of bins is treated as our operating context. We also present as a case study a real application in sentiment analysis where these different operating contexts are given. As mentioned, for some applications a '*Good* $\succ$ *Bad*' resolution can be good enough whereas for some others, an "*Excellent* $\succ$ *Good* $\succ$ *Fair* $\succ$ *Poor* $\succ$ *Very bad*" granularity is better. The number of categories or bins is only known during deployment. A very similar approach in [5] presents the performance of different methodologies as the number of sentiment categories increases. In our work, we compare different regression techniques and a hybrid of the best produced results for the range of contexts over a set of regression problems. In order to measure the results of the different methodologies, we have used the *macro-discrepancy measure* (mD), namely the mean of the absolute rank errors across classes weighting all classes equally, which has been proposed as robust alternative for imbalanced datasets [2, 13].

For the sake of brevity, the core contributions of this paper are outlined as follows:

- We show that the information about the number of possible bin contexts we have to deal with during the deployment phase can be exploited during a model validation stage to analyse which learning technique performs better for each number of bins. For our study, we use well-known regression techniques, showing graphical results of their performance with respect to the number of context bins. The analysis of these graphs allows us to determine for a given problem which model can be discarded for any number of bins or which model is the best one given a fixed context.
- We also propose a new hybrid model based on model dominance (hereafter *Hybrid* model) that performs reasonably well in any dataset and for all the contexts.
- We present a sentiment analysis case study using a real-world dataset [5] of reviews labelled with scores on a 91 point scale (1.0 to 10.0 with a 0.1 resolution) and compare the performance of the *Hybrid* model with several machine learning algorithms.

The rest of the paper is organised as follows. Section 2 describes our notion of context and the process of how the models are adapted to each context. Section 3 explains the *Hybrid* model proposed on this work and defines the error measure that will be used in the experiments. Section 4 shows how different regression techniques and the proposed *Hybrid* model behave using a real-world dataset. Section 5 presents the techniques and datasets used for the experimental analysis of our approach. After that, results for each approach are analysed, discussing whether there are general patterns about which method can be best for some contexts. Section 6 closes the paper with a discussion about related works and which take-away messages we would like to focus on. Future work is also discussed in this section.

2 Context as number of bins in ordinal regression problems

This section describes a process for facing the problem of adapting a regression model to several contexts, being each context defined by the number of bins in which we want to classify the deployment data. These contexts, which we must adapt to, are called the *operating conditions* of the problem. Figure 1 summarises the process of discretisation of the regression values into bins to adapt models to contexts. Note also that this flowchart only shows the discretisation procedure, not the entire experimental setting described in Sect.5.

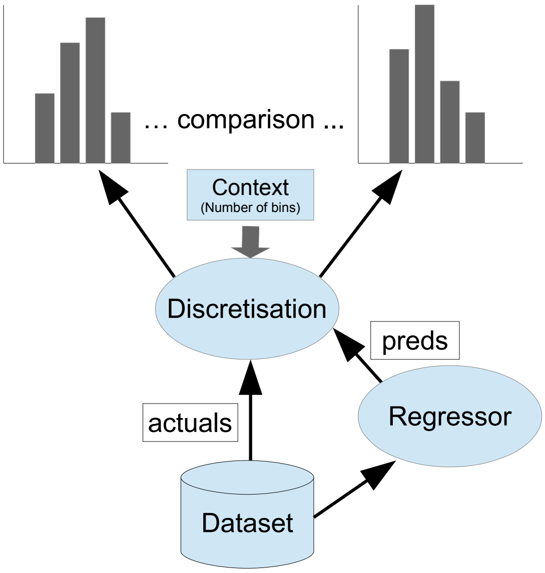

Fig. 1. Model context adaptation procedure flowchart. Histograms on the top show how the samples are distributed in bins according to their true or predicted values and the number of bins of the context.

From a regression dataset, a regression technique is trained, producing its predictions (*preds*) for a test set. These predictions together with the true values (*actuals*) are discretised once the number of bins Q is known. The discretisation process divides the range of values of the *actuals* into N equal-sized bins. The width b_{width} of the interval of values that constitutes each of the bins is worked out as $b_{width} = \frac{d_{max}-d_{min}}{Q}$, where $d_{max} = max(actuals)$ and $d_{min} = min(actuals)$.

3 A hybrid model and evaluation metric for ordinal classification

In this section, the proposed hybrid model for ordinal classification and the evaluation measure we are going to use are presented. Our proposal assumes that a range of regression methods are applied to a certain dataset and we wonder whether we can determine for which number of bins each method performs better during training. Then, we propose a simple yet very useful method, called *Hybrid* model. The idea is to see the different dominance regions where a method outperforms the others, and always use the best one for the context in hand. Note that this is different of an ensemble in that the whole collection of models would be used for any context. The procedure to obtain the *Hybrid* model is summarised as follows:

1. Given a regression dataset, we divide it into training and validation subsets and use the former for training several models.
2. Different number of bins according to the operating context are added to the validation data following an equal-width setting for the bins. Models are then evaluated on the validation subset for each bin.
3. A context plot is drawn with all the models. The dominant method for each number of bins is identified.
4. In deployment time and with different data, the operating context is mapped to the number of bins we are interested in. The best model for that number of bins is applied.

The performance of an ordinal classification approach is often measured as a multiclass classification problem for which there is an inherent order between the classes. Unfortunately, there is no clear measure preferred in this research field [4]. In our work we use the *macro-discrepancy measure*[1] (mD) for measuring the different methodologies:

$$mD = \frac{1}{Q}\sum_{j=1}^{Q} mD_j = \frac{1}{Q}\sum_{j=1}^{Q}\frac{1}{n_j}\sum_{i=1}^{n_j}|O(y_i) - O(\hat{y}_i)| \tag{1}$$

Function O gives the rank of a sample, where y_i and $\hat{y}_i$ are the true and predicted values respectively. mD works out mD_j for each group of samples j within the

[1] This measure is called AMAE (*average MAE*) in the literature [2, 13]. We use a different name because MAE is very confusing as it can be understood as the mean absolute error of the regression model, which is not.

same class or bin (grouped by true labels), for Q bins. n_j is the number of samples in class j. mD ranges from 0 (when the rank of the samples is always right) to $Q-1$ (when the utmost difference between ranks is given) and evaluates the mean of the absolute rank errors across classes. The use of a macro-average (or stratified average) instead of a micro-average is because it is said to be more robust for imbalanced datasets [2, 13].

4 Case study

This work got inspiration from [5], where the authors addressed the problem of considering different number of sentiment categories in a real-valued sentiment analysis approach. In this section we show how several regression techniques and our proposed *Hybrid* model perform on the same dataset they used in their work. We use the following regression methods for this case study (all of them will also be used in the experimental section): `RegrTree` (a regression tree), `LinearRegression` (a linear regression), `SMOreg` (a support vector machine) and `IBk` (a k nearest neighbour). Thus, Figure 2 shows these performances in terms of the mD measure and for a range of bins between 2 and 12. As it can be seen, almost all the methods perform similarly, being only the `LinearRegression` model the one that obtains quite poor performance.

Considering Figure 2, we define the measure APC as the area under the curve constructed by plotting the performance measure mD for the range of the number of bins analysed (2-12). Table 1 summarises the previous graph with the APC values for each of the studied methods. Although quite similar, the regression SVM (`SMOreg`) shows the best performance in general.

	RegrTree	LinearRegression	SMOreg	IBk	Hybrid
DB_SentAnalysis	17.65	22.67	**16.97**	17.42	17.09
Rank	4	5	1	3	2

Table 1. APC measure of the studied methods for the sentiment analysis dataset [5].

5 Experimental section

In order to study different procedures of selecting models when the number of bins in test data is unknown, we performed a set of experiments over 25 regression datasets from several repositories [3, 14] and 5 regression techniques. Datasets are shown in Table 2.

For each dataset, we split the data into 3 different sets: Training (50%), Validation (25%) and Test (25%). In order to reduce the variance of results, we repeat 5 times the partition of data. We use the training data for learning the models and the validation data to estimate the performance of the models

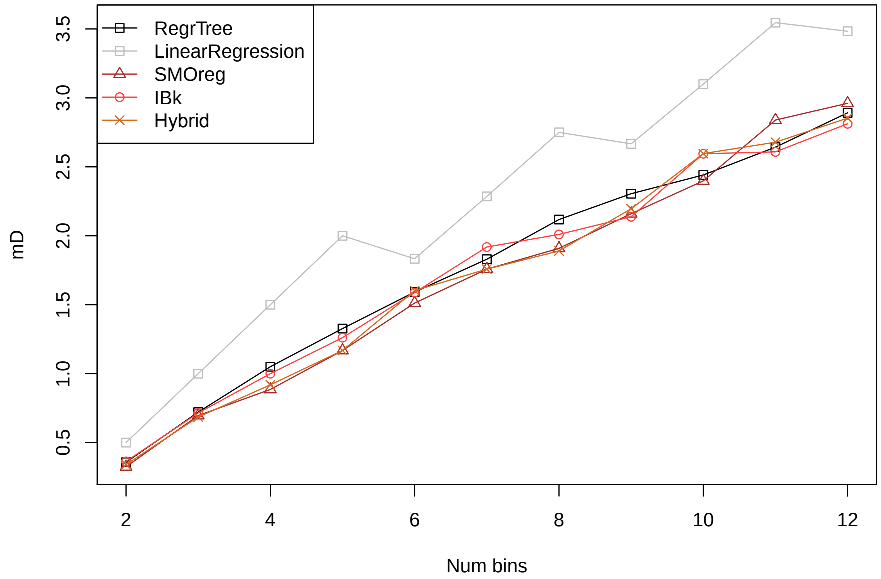

Fig. 2. Performance of the regression techniques for each number of sentiment categories (bins) for the Sentiment Analysis dataset.

when the number of bins changes. Then, we use the remaining test set as the deployment dataset.

For the learning techniques, we use the *RWeka* [11] library from *R*. Concretely, as we have mentioned in the previous section, for regression problems we use the following techniques: A decision tree for regression (`RegrTree`), a Linear Regression (`LinearRegression`), a Support Vector Machines method for regression (`SMOreg`) and kNN (`IBk`). We also include the `ZeroR` method as a baseline. This method returns for all examples the mean of the train dataset. After discretisation, this model will predict all examples in the same bin.

In addition to the previous well-known learning techniques, we have also considered a *Hybrid* model (`Hybrid`) as the approach that, for each number of bins, selects the best performance model in the validation dataset (i.e., the algorithm with lowest value in the context plot, the hybrid method explained in Section 3). The question to be answered is the performance of this *Hybrid* model approach.

The plots in Figure 3 show some examples of how the different methodologies behave in terms of mD (y-axis) according to the number of bins (x-axis). The two graphs on the top are situations where `Hybrid` performs fine (*plastic* and

dataset	NumInst	NumAtt	Ave	Sd	Max	Min
01airfoil_self_noise	1503	6	124.836	6.899	140.987	103.380
02ENB2012_data_cooling	768	9	24.588	9.513	48.030	10.900
03ENB2012_data_heating	768	9	22.307	10.090	43.100	6.010
04CCPP	9568	5	454.365	17.067	495.760	420.260
05Concrete_Data	1030	9	35.818	16.706	82.600	2.330
06autoMpg	398	8	23.515	7.816	46.600	9.000
07housing	506	14	22.533	9.197	50.000	5.000
08abalone	4177	9	9.934	3.224	29.000	1.000
09yacht_hydrodynamics	308	7	10.495	15.160	62.420	0.010
10winequality-white	4898	12	5.878	0.886	9.000	3.000
11winequality-red	1599	12	5.636	0.808	8.000	3.000
12solar-flare_1	323	11	0.291	0.769	6.000	0.000
13diabetes	43	3	4.747	0.721	6.600	3.000
14dee	365	7	2.971	0.966	5.119	0.766
15plastic	1650	3	15.000	3.417	20.000	10.000
16treasury	1049	16	7.522	3.377	20.760	3.020
17wankara	321	10	48.921	14.976	81.600	16.200
18wizmir	1461	10	61.508	14.376	89.900	29.400
19cpu_small	8192	13	83.969	18.402	99.000	0.000
20auto_price	159	16	11445.7	5877.9	35056.0	5118.0
21pyrim	74	28	0.659	0.128	0.900	0.100
22wisconsin	194	33	46.938	34.524	125.000	1.000
23delta_ailerons	7129	6	-0.000	0.000	0.002	-0.002
24delta_elevators	9517	7	-0.000	0.002	0.013	-0.014
25triazines	186	61	0.652	0.158	0.900	0.100

Table 2. Information about the regression datasets used in the experiments. From left to right: number of instances, number of attributes, average, standard deviation, maximum and minimum.

delta_elevators datasets). On the other hand, two graphs on the bottom are situations where the `Hybrid` shows a worse behaviour (*diabetes* and *auto_price* datasets).

In order to compare all datasets and have an overall view, we have done two things. Analyse ranks and average them, and analyse the mD values and average them. As these are not commensurate, we have use the baseline method `ZeroR` to normalise the mD values. Namely, all the error measures obtained by the techniques will be divided by the `ZeroR` error value, facilitating thereby comparisons among the methods.

Table 3 shows the average results of all executions for each learning method. As expected, the *Hybrid* model performs quite well although the regression tree algorithm (`RegrTree`) obtains the best ranking (last row in the table). `Hybrid` always shows a very similar performance to the best algorithm. Because of this, for the aggregate values (APC), the *Hybrid* model is better.

In non-parametric statistics, a very popular procedure for testing the differences between more than two related samples is the Friedman test [9]. It ranks

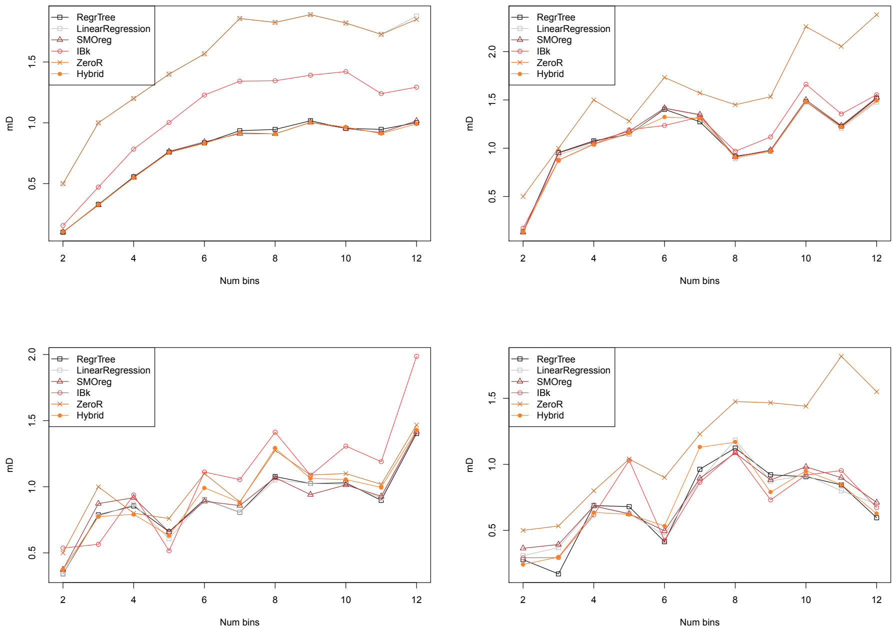

Fig. 3. Four example graphs from datasets. From top to bottom, left to right, *plastic*, *delta_elevators*, *diabetes* and *auto_price* datasets.

the number of algorithms (in our case 5, columns in Table 3) for each of the datasets (in our case 25, rows in Table 3) independently according to the performance obtained, that is, from 1 (best) to 5 (worst). Table 3 shows in its bottom row the results of the Friedman ranking where the Regression Tree algorithm has obtained the best rank, followed by our proposal.

However, this counts any difference, even the slightest one, for the ranking. A very well-known and powerful non-parametric statistical test in detecting the existence of significant differences between pairs of methods is the signed-rank Wilcoxon non-parametric procedure. Thus, in order to give another view that supports our results, the Wilcoxon test has also been performed[2]. Table 4 summarises the result of the Wilcoxon test; where a full bullet symbol (•) means that the method in the row significantly improves the method in the column whereas an empty bullet symbol (◦) means that the method in the column significantly improves the method in the row. Upper diagonal for a level of significance $\alpha = 0.9$. Lower diagonal for a level of significance a $\alpha = 0.95$. An empty cell means that no significant improvement can be established between the two methods. As it can be seen, both `RegrTree` and our proposed `Hybrid`

[2] Both non-parametric tests have been performed using software in [1].

DATASETS	RegrTree	LinearRegression	SMOreg	IBk	Hybrid
01airfoil_self_noise	0.5093	0.6682	0.6410	**0.4101**	0.4142
02ENB2012_data_cooling	**0.5230**	0.6151	0.6328	0.6128	0.5296
03ENB2012_data_heating	**0.2122**	0.4645	0.4895	0.4852	**0.2122**
04CCPP	0.4055	0.4713	0.4522	**0.3051**	**0.3051**
05Concrete_Data	**0.5683**	0.7470	0.7442	0.5817	0.5715
06autoMpg	0.5775	**0.5764**	0.5876	0.6131	0.5985
07housing	**0.4538**	0.5900	0.6000	0.6064	0.4959
08abalone	**0.7324**	0.7532	0.7740	0.7409	0.7349
09yacht_hydrodynamics	**0.3866**	0.6607	0.8035	0.9639	0.3904
10winequality-white	0.9093	0.9253	0.9199	**0.7830**	0.7963
11winequality-red	0.8803	0.8867	0.8931	**0.8425**	0.8616
12solar-flare_1	0.8644	**0.8570**	0.8598	0.8956	0.8655
13diabetes_numeric	0.8904	**0.8842**	0.9035	1.0433	0.9368
14dee	0.5413	**0.5326**	0.5536	0.6175	0.5533
15plastic	0.5064	1.0008	0.5008	0.7084	**0.4994**
16treasury	**0.1690**	0.2029	0.2252	0.1783	0.1817
17wankara	**0.1650**	0.2089	0.2148	0.3514	0.1817
18wizmir	**0.1045**	0.1087	0.1087	0.2574	0.1057
19cpu_small	**0.1917**	0.4185	0.3594	0.2322	**0.1917**
20auto_price	**0.6093**	0.6228	0.6377	0.6316	0.6317
21pyrim	0.7831	0.6879	0.6525	**0.4882**	0.6017
22wisconsin	**1.0126**	1.0552	1.0226	1.1940	1.0488
23delta_ailerons	**0.6932**	0.7453	0.7757	0.7314	0.7091
24delta_elevators	0.7125	0.7093	0.7211	0.7351	**0.7014**
25triazines	0.8692	0.8872	**0.8588**	0.9007	0.8643
Avg.AccumPercent.	0.5708	0.6512	0.6373	0.6364	**0.5593**
Avg.ranks (Ftest)	**1.94**	3.44	3.76	3.52	2.34

Table 3. *APC* for regression datasets. Values normalised by the ZeroR method. The best performance on each dataset is written in bold type. Ftest stands for the average rankings of Friedman test.

significantly improve the rest of the methods for both levels of significance, being no significant difference between them.

6 Discussion

As mentioned, this work follows [5], where the authors addressed the problem of considering different number of sentiment categories in a real-valued sentiment analysis approach. In their work, authors compared the performance of several techniques when the number of classes/bins (sentiment categories) increases. We have used their study as a running example for a real application problem, extending their approach with a new hybrid method that combines the performances of several regression techniques and by applying this methodology to a wide range of regression datasets.

	(1)	(2)	(3)	(4)	(5)	(6)
`RegrTree` (1)	-	●	●	●	●	
`LinearRegression` (2)	○	-			●	○
`SMOreg` (3)	○		-		●	○
`IBk` (4)	○			-	●	○
`ZeroR` (5)	○	○	○	○	-	○
`Hybrid` (6)		●	●	●	●	-

Table 4. Summary of the Wilcoxon test. $\bullet=$ the method in the row improves the method in the column. $\circ=$ the method in the column improves the method in the row. Upper diagonal of level significance $\alpha = 0.9$, Lower diagonal level of significance $\alpha = 0.95$.

The use of context plots to determine the best model when the context changes has been also studied in [7], where authors faced (traditional) regression and classification problems when the context (level of noise in the data) is known. In [10], the authors proposed two approaches to deal with the problem of transforming a regression problem into a binarised classification one. The basic idea in these approaches was that, for many applications, we are interested in knowing whether a prediction is above or below a given cutoff (binary classification). This idea is somehow related to our work, since we also use the predictions of the regressors for multiclass ordinal classification (with a number of classes ≥ 2), being our aim to determine the best method for the context in the deployment step.

Although preliminary, we have shown in this work that the information about the number of possible bin contexts we have to deal with during the deployment phase can be exploited during a model validation stage to analyse which learning technique performs better for each number of bins.

This work also suggests some avenues for future work. We are especially interested in exploring this approach for a equal-frequency binning, dividing the range of observed values into a given number of intervals so that the number of instances in each interval remains approximately constant (in training and deployment). There is also room for improvement in (i) the discretisation process since more sophisticated techniques can be applied when translating the outputs into a discrete class value [8] and (ii) the comparison by means of adding some other interesting baselines (e.g. training a different multi-class classifier for each number of classes).

Acknowledgements

Authors would like to thank Professor Dan Ventura for his help with the Sentiment Analysis dataset. This work was supported by the Spanish MINECO under grant TIN 2013-45732-C4-1-P, and the REFRAME project, granted by the European Coordinated Research on Long-term Challenges in Information and Communication Sciences Technologies ERA-Net (CHIST-ERA), and funded by

MINECO in Spain (PCIN-2013-037) and by Generalitat Valenciana PROMETEOII2015/013.

References

1. Alcalá-Fdez, J., Sánchez, L., García, S., del Jesus, M., Ventura, S., Garrell, J., Otero, J., Romero, C., Bacardit, J., Rivas, V., Fernández, J., Herrera, F.: Keel: a software tool to assess evolutionary algorithms for data mining problems. Soft Computing 13(3), 307–318 (2009)
2. Baccianella, S., Esuli, A., Sebastiani, F.: Evaluation measures for ordinal regression. In: Intelligent Systems Design and Applications, ISDA'09. Ninth International Conference on. pp. 283–287 (Nov 2009)
3. Bache, K., Lichman, M.: UCI machine learning repository (2013), `http://archive.ics.uci.edu/ml`
4. Cardoso, J.S., Sousa, R.: Measuring the performance of ordinal classification. International Journal of Pattern Recognition and Artificial Intelligence 25(08), 1173–1195 (2011)
5. Drake, A., Ringger, E., Ventura, D.: Sentiment regression: Using real-valued scores to summarize overall document sentiment. In: Semantic Computing, IEEE International Conference on. pp. 152–157 (Aug 2008)
6. Feldman, R.: Techniques and applications for sentiment analysis. Commun. ACM 56(4), 82–89 (2013)
7. Ferri, C., Hernández-Orallo, J., Martínez-Usó, A., Ramírez-Quintana, M.: Identifying dominant models when the noise context is known. In: Intelligent Systems Design and Applications, ISDA'09. Ninth International Conference on (2014)
8. Frank, E., Hall, M.: A simple approach to ordinal classification. In: De Raedt, L., Flach, P. (eds.) Machine Learning: ECML 2001, Lecture Notes in Computer Science, vol. 2167, pp. 145–156. Springer Berlin Heidelberg (2001)
9. Friedman, M.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association 32(200) (1937)
10. Hernández-Orallo, J., Ferri, C., Lachiche, N., Martínez-Usó, A., Ramírez-Quintana, M.: Binarised regression tasks: Methods and evaluation metrics. Submitted to Data Mining and Knowledge Discovery
11. Hornik, K., Buchta, C., Zeileis, A.: Open-source machine learning: R meets Weka. Computational Statistics 24(2), 225–232 (2009)
12. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal 5(4), 1093–1113 (2014)
13. Sánchez-Monedero, J., Gutiérrez, Antonio, P., Tino, P., Hervás-Martínez, C.: Exploitation of Pairwise Class Distances for Ordinal Classification. Neural Computation 25(9), MIT Press (2013)
14. Torgo, L.: Regression datasets (2001), `http://www.dcc.fc.up.pt/~ltorgo/Regression/DataSets.html`