Combining Absolute and Relative Information with Frequency Distributions for Ordinal Classification

Mengzi Tang[1(✉)], Raúl Pérez-Fernández[1,2], and Bernard De Baets[1]

[1] KERMIT, Department of Data Analysis and Mathematical Modelling, Ghent University, Coupure links 653, 9000 Ghent, Belgium
{mengzi.tang,raul.perezfernandez,bernard.debaets}@ugent.be
[2] Department of Statistics and O.R. and Mathematics Didactics, University of Oviedo, C/ Federico García Lorca 18, 3307 Oviedo, Spain

Abstract. A large amount of labelled data (absolute information) is usually needed for an ordinal classifier to attain a good performance. As shown in a recent paper by the present authors, the lack of a large amount of absolute information can be overcome by additionally considering some side information in the form of relative information, thus augmenting the method of nearest neighbors. In this paper, we adapt the method of nearest neighbors for dealing with a specific type of relative information: frequency distributions of pairwise comparisons (rather than a single pairwise comparison). We test the proposed method on some classical machine learning datasets and demonstrate its effectiveness.

Keywords: Ordinal classification · Nearest neighbors · Absolute information · Relative information · Frequency distributions

1 Introduction

Typically for ordinal classification there is only absolute information available, i.e., examples with an associated class label of a fixed ordinal scale. Unfortunately, in real-life applications, it is often the case that the amount of absolute information available is limited, thus largely impacting the performance of an ordinal classifier. Fortunately, different types of side information can be additionally collected and make up for the limitation regarding the little amount of absolute information available [1,2]. A popular type of such side information is relative information, i.e., couples of examples without an explicitly given class label but with an order relation between them.

Interestingly, in real-life applications, relative information with frequency distributions arises quite commonly. For instance, the emergence of online transaction platforms such as Amazon Mechanical Turk offers some possibilities to distribute evaluation tasks to consumers and collect a large amount of relative information. However, the order relations from relative information are usually

M.-J. Lesot et al. (Eds.): IPMU 2020, CCIS 1238, pp. 594–602, 2020.
https://doi.org/10.1007/978-3-030-50143-3_47

less informative than the class labels from absolute information. It is also quite common for customers to have contradictory order relations for the same couple of examples. Due to these facts, it is better to consult several customers for collecting the preference among two examples, thus obtaining for each couple of examples a frequency distribution of order relations (hereinafter referred to as relative information with frequency distributions). Hence, how to combine a small amount of absolute information and a large amount of relative information with frequency distributions becomes our main goal.

Some related works [3,4] have shown the effectiveness of fusing absolute and relative information. In the field of ordinal classification, for example, Sader et al. [5] proposed an ordinal classification method for combining both absolute and relative information to perform prediction tasks. This method needs to learn many parameters for solving a constrained convex optimization problem. In a similar direction, our previous work [6] incorporated both types of information into the method of nearest neighbors, and proposed an augmented method for ordinal classification that is non-parametric and easy to explain. However, this method was designed to deal with just one order relation for each couple of examples and not with a frequency distribution of order relations. An immediate extension of this method to the latter setting reduces the study of the nearest couples of examples to just the nearest couple of examples, thus impacting its overall performance. To properly address our problem setting, where there is a small amount of absolute information and a large amount of relative information with frequency distributions available, we propose a method to incorporate both types of information into the method of nearest neighbors for ordinal classification on the basis of our previous work [6].

The remainder of this paper is organized as follows. In Sect. 2, we formulate our problem. We propose our method in Sect. 3. In Sect. 4, we perform experiments and analyze the performance. Some conclusions are presented in Sect. 5.

2 Problem Setting

The available data is composed of absolute and relative information. We denote the input examples by $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. The input examples $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})$ belong to the input space $\mathcal{X} \subseteq \mathbb{R}^d$ and their corresponding class labels y_i belong to the output space $\mathcal{Y} = \{C_1, C_2, \ldots, C_r\}$, where the class labels are supposed to be ordered as follows: $C_1 \prec C_2 \prec \ldots \prec C_r$. Absolute information is collected in a set $\mathcal{A} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$.

Although for some examples there is no explicitly given class label, it is still possible to have some side information in the form of relative information. Relative information is typically expressed for a set of couples of examples $\mathcal{C} = \{(\mathbf{a}^i, \mathbf{b}^i), \ldots, (\mathbf{a}^m, \mathbf{b}^m)\} \in \mathcal{X}^2$. With each couple $(\mathbf{a}^i, \mathbf{b}^i)$, a frequency distribution (α^i, β^i) is associated, α^i representing the proportion of times that $\mathbf{a}^i$ is preferred to $\mathbf{b}^i$ and β^i representing the proportion of times that $\mathbf{b}^i$ is preferred to $\mathbf{a}^i$. Obviously, $\alpha^i + \beta^i = 1$. Relative information is collected in a set $\mathcal{R} = \{((\mathbf{a}^1, \mathbf{b}^1), (\alpha^1, \beta^1)), ((\mathbf{a}^2, \mathbf{b}^2), (\alpha^2, \beta^2)), \ldots, ((\mathbf{a}^m, \mathbf{b}^m), (\alpha^m, \beta^m))\}$.

If $((\mathbf{a}^i, \mathbf{b}^i), (\alpha^i, \beta^i))$ belongs to $\mathcal{R}$, then $((\mathbf{b}^i, \mathbf{a}^i), (\beta^i, \alpha^i))$ is supposed also to belong to $\mathcal{R}$. Note that here we do not consider the case in which $\mathbf{a}^p$ and $\mathbf{b}^p$ are equally preferred. The main characteristic of our problem is that the amount of absolute information is typically smaller than the amount of relative information, i.e., $n \ll m$.

3 Proposed Method

3.1 Existing Method: Fusing Absolute and Relative Information for Augmenting the Method of Nearest Neighbors

In this subsection, we recall the method proposed in our previous work [6]. Firstly, according to a fixed distance metric d, we find the k nearest neighbor examples $\mathcal{D}_k = \{\mathbf{x}_{i_j}\}_{j=1}^k$ of the test example $\mathbf{x}^*$. We see each couple $(\mathbf{x}^*, \mathbf{x}_{i_j})$ as a new object and look for the ℓ nearest neighbor couples $\mathcal{C}_\ell^j = \{(\mathbf{a}_q^j, \mathbf{b}_q^j)\}_{q=1}^\ell$ of this new object $(\mathbf{x}^*, \mathbf{x}_{i_j})$. For this process, we compute the distance between couples according to the product distance metric (see [7], page 83, with p=1), which is defined as

$$d_*((\mathbf{u}, \mathbf{v}), (\mathbf{w}, \mathbf{t})) = d(\mathbf{u}, \mathbf{w}) + d(\mathbf{v}, \mathbf{t}). \tag{1}$$

Secondly, we rely on the assumption that a couple and its nearest neighbor couples have similar order relations. More in detail, for the new object $(\mathbf{x}^*, \mathbf{x}_{i_j})$, we focus on the nearest neighbor couple and get its corresponding order relation. For instance, if the nearest neighbor couple of $(\mathbf{x}^*, \mathbf{x}_{i_j})$ is $(\mathbf{a}_1^j, \mathbf{b}_1^j)$ and its given order relation is $\mathbf{a}_1^j \succ \mathbf{b}_1^j$, then we assume the same order relation $\mathbf{x}^* \succ \mathbf{x}_{i_j}$ for $(\mathbf{x}^*, \mathbf{x}_{i_j})$. Since the class label of $\mathbf{x}_{i_j}$ is known to be, for instance, C_{c_j}, the class label of $\mathbf{x}^*$ is expected to be at least C_{c_j}. The same applies to the other $\ell - 1$ neighbor couples. For each among these ℓ relations, we obtain an interval of potential class labels for $\mathbf{x}^*$. We denote this interval as $I_{jq} = [l_{I_{jq}}, r_{I_{jq}}]$, where $j \in \{1, \ldots, k\}$ and $q \in \{1, \ldots, \ell\}$. For instance, if the given class label of $\mathbf{x}_{i_j}$ is C_{c_j} and we obtain that the relation for the couple $(\mathbf{x}^*, \mathbf{x}_{i_j})$ is $\mathbf{x}^* \succ \mathbf{x}_{i_j}$ according to its q-th nearest neighbor couple, then the interval of possible values of y^* is $I_{jq} = [l_{I_{jq}}, r_{I_{jq}}] = [C_{c_j}, C_r]$. Similarly, if the relation is $\mathbf{x}^* \prec \mathbf{x}_{i_j}$, then the interval of possible values of y^* is $I_{jq} = [l_{I_{jq}}, r_{I_{jq}}] = [C_1, C_{c_j}]$.

Finally, we denote by $\mathbf{I} = (I_{jq})_{j \in \{1,\ldots,k\}, q \in \{1,\ldots,\ell\}}$ the list of all the obtained intervals. We consider the penalty function associated with the median for intervals (see, for instance, Beliakov et al. [8]):

$$P(\mathbf{I}, y) = \sum_{j=1}^{k} \sum_{q=1}^{\ell} (|l_{I_{jq}} - y| + |r_{I_{jq}} - y|), \tag{2}$$

where $|C_i - C_j|$ denotes the L_1-distance between two class labels C_i and C_j. Note that the L_1-distance metric treats all class labels of the ordinal scale as if they were equidistant, something that is not always advisable depending on the

nature of $\mathcal{Y}$. The class label y^* of $\mathbf{x}^*$ is then determined using the corresponding penalty-based (aggregation) function:

$$y^* = f(\mathbf{y}^*) = \arg\min_{y \in \mathcal{Y}} P(\mathbf{I}, y). \tag{3}$$

3.2 New Method: Combining Absolute and Relative Information with Frequency Distributions

The method above only focuses on how to deal with couples provided with only one order relation. However, in our problem setting, we have relative information with frequency distributions. More specifically, we have a frequency distribution of order relations for each $(\mathbf{a}^j, \mathbf{b}^j)$. We use (α^j, β^j) to characterize this frequency distribution. In the following, we explain how to deal with such information.

Firstly, we repeat the process above to look for the k nearest neighbor examples of the test example $\mathbf{x}^*$. We see each couple $(\mathbf{x}^*, \mathbf{x}_{i_j})$ as a new object, search for the ℓ nearest neighbor couples, and get their frequency distributions of order relations. We rely on the same assumption above that a couple and its nearest neighbor couples have similar order relations. More in detail, if the nearest neighbor couple of $(\mathbf{x}^*, \mathbf{x}_{i_j})$ is $(\mathbf{a}_q^j, \mathbf{b}_q^j)$ and its frequency distribution is (α_q^j, β_q^j), which implies that a proportion α_q^j of times the order relation of $(\mathbf{a}_q^j, \mathbf{b}_q^j)$ is $\mathbf{a}_q^j \succ \mathbf{b}_q^j$ and a proportion β_q^j of times the order relation of $(\mathbf{a}_q^j, \mathbf{b}_q^j)$ is $\mathbf{a}_q^j \prec \mathbf{b}_q^j$, then in case the couple $(\mathbf{x}^*, \mathbf{x}_{i_j})$ needs to be labelled, we would expect that a proportion α_q^j of times the order relation of $(\mathbf{x}^*, \mathbf{x}_{i_j})$ is $\mathbf{x}^* \succ \mathbf{x}_{i_j}$ and a proportion β_q^j of times the order relation of $(\mathbf{x}^*, \mathbf{x}_{i_j})$ is $\mathbf{x}^* \prec \mathbf{x}_{i_j}$.

For the new couple $(\mathbf{x}^*, \mathbf{x}_{i_j})$ in which the given class label of $\mathbf{x}_{i_j}$ is C_{c_j}, we get α_q^j times the interval $[C_{c_j}, C_r]$ and β_q^j times the interval $[C_1, C_{c_j}]$ for the potential class label y^* of $\mathbf{x}^*$. We repeat this process for the other $\ell - 1$ nearest neighbor couples. Exploiting all k nearest neighbors and ℓ nearest neighbor couples, we get a list of intervals of potential class labels for $\mathbf{x}^*$. We denote by $\mathbf{I}$ the list of all gathered intervals.

Finally, differently to the previous section, here we do not use the notation $I_{jq} = [l_{I_{jq}}, r_{I_{jq}}]$ to represent the interval. More specifically, we now have a proportion α_q^j of times the interval $[C_{c_j}, C_r]$ and a proportion β_q^j of times the interval $[C_1, C_{c_j}]$ for each nearest neighbor couple of $(\mathbf{x}^*, \mathbf{x}_{i_j})$. Thus, the penalty function associated with the median reads as follows:

$$\begin{aligned} P(\mathbf{I}, y) &= \sum_{j=1}^{k} \sum_{q=1}^{\ell} \beta_q^j (|C_1 - y| + |C_{c_j} - y|) + \alpha_q^j (|C_{c_j} - y| + |C_r - y|) \\ &= \sum_{j=1}^{k} \sum_{q=1}^{\ell} (\beta_q^j |C_1 - y| + |C_{c_j} - y| + \alpha_q^j |C_r - y|), \end{aligned} \tag{4}$$

where $(\mathbf{a}_q^j, \mathbf{b}_q^j)$ is the q-th nearest neighbor couple of the couple $(\mathbf{x}^*, \mathbf{x}_{i_j})$, (α_q^j, β_q^j) is the corresponding frequency distribution and the given class label of the j-th

Table 1. Description of the benchmark datasets.

Dataset	#Examples	#Features	#Classes
Real ordinal classification datasets			
Tae (*TA*)	151	54	3
Automobile (*AU*)	205	26	6
Balance-scale (*BS*)	625	4	3
Eucalyptus (*EU*)	736	91	5
Red-wine (*RW*)	1599	12	6
Car (*CA*)	1728	21	4
Discretized regression datasets			
*Housing*5 (*HO*5)	506	14	5
*Abalone*5 (*AB*5)	4177	11	5
*Bank*1-5 (*BA*1-5)	8192	8	5
*Bank*2-5 (*BA*2-5)	8192	32	5
*Computer*1-5 (*CO*1-5)	8192	12	5
*Computer*2-5 (*CO*2-5)	8192	21	5
*Housing*10 (*HO*10)	506	14	10
*Abalone*10 (*AB*10)	4177	11	10
*Bank*1-10 (*BA*1-10)	8192	8	10
*Bank*2-10 (*BA*2-10)	8192	32	10
*Computer*1-10 (*CO*1-10)	8192	12	10
*Computer*2-10 (*CO*2-10)	8192	21	10

nearest neighbor $\mathbf{x}_{i_j}$ of $\mathbf{x}^*$ is C_{c_j}. The class label y^* of $\mathbf{x}^*$ is then determined using the corresponding penalty-based (aggregation) function:

$$y^* = f(\mathbf{y}^*) = \arg\min_{y \in \mathcal{Y}} P(\mathbf{I}, y). \tag{5}$$

4 Experiments

4.1 Datasets

We perform the experiments on some classical machine learning datasets from some typical repositories [9–11]. The detailed characteristics of these datasets can be found in Table 1, including the number of examples, features and classes. All the features have been properly normalized (by making all the features to have zero mean and unit standard deviation) to avoid the impact of the scale of features. Note that the datasets do not contain relative information with frequency

distributions. Based on a similar generation process of relative information as the one described in [6], we generate couples with frequency distributions of order relations.

More in detail, when comparing two examples for generating a couple $(\mathbf{a}^i, \mathbf{b}^i)$, we randomly sample α^i or β^i from a uniform distribution. For example, if the real order relation of these two examples is $\mathbf{a}^i \succ \mathbf{b}^i$, then we sample α^i from a uniform distribution on [0.5, 1] and set $\beta^i = 1 - \alpha^i$. Similarly, if the real order relation of these two examples is $\mathbf{a}^i \prec \mathbf{b}^i$, then we sample β^i from a uniform distribution on [0.5, 1] and set $\alpha^i = 1 - \beta^i$. Thus, we generate a couple $((\mathbf{a}^i, \mathbf{b}^i), (\alpha^i, \beta^i))$.

To test our method, we construct two different datasets for each original dataset. Based on a similar generation process as in our previous work [6], we fix 10% of the data that will be shared by both datasets for testing. The remaining 90% is used for generating the data for training. We keep 5% of the remaining 90% as absolute information. We use the remaining 95% for generating relative information following the aforementioned description. *Dataset 1* includes just absolute information. *Dataset 2* not only includes the same absolute information as *Dataset 1*, but also includes relative information with frequency distributions. By comparing the performance on these two datasets, we test the impact of incorporating relative information with frequency distributions.

4.2 Performance Measures

We use the three most common performance measures to evaluate ordinal classification models [13,14]: the Mean Zero-one Error (MZE), the Mean Absolute Error (MAE) and the C-index.

The MZE describes the error rate of the classifier and is computed as

$$\text{MZE} = \frac{1}{T} \sum_{i=1}^{T} \delta(y_i^* \neq y_i) = 1 - \text{Acc}, \tag{6}$$

where T is the number of test examples, y_i is the real class label and y_i^* is the predicted class label. Acc is the accuracy of the classifier. The value of MZE ranges from 0 to 1. It describes the global performance, but it neglects the relations among the class labels.

The MAE is the average absolute error between y_i and y_i^*. If the class labels are represented by numbers, the MAE is computed as:

$$\text{MAE} = \frac{1}{T} \sum_{i=1}^{T} |y_i - y_i^*|. \tag{7}$$

The value of MAE ranges from 0 to $r - 1$ (maximum absolute error between classes). Because the real distances among the class labels are unknown, the numerical representation of the class labels has a big impact on the MAE performance.

In order to avoid this impact, one could consider the relations between the real class label and the predicted class label. Here we use the concordance index

Table 2. Performances on the two newly constructed datasets for each original dataset. The best results are highlighted in boldface.

Dataset	MZE		MAE		1 – C-index	
	Dataset 1	Dataset 2	Dataset 1	Dataset 2	Dataset 1	Dataset 2
TA	0.5959	**0.5839**	0.7664	**0.6889**	0.4549	**0.3963**
AU	0.6047	**0.5395**	0.9410	**0.7522**	0.3804	**0.2749**
BS	0.2366	**0.2080**	0.3549	**0.3039**	0.1817	**0.1547**
EU	0.6383	**0.5693**	1.0267	**0.8329**	0.3275	**0.2539**
WR	0.4997	**0.4914**	0.5747	**0.5608**	0.3338	**0.3180**
CA	0.2269	**0.2205**	0.2778	**0.2744**	0.2440	**0.1976**
HO5	0.5496	**0.5089**	0.7456	**0.6837**	0.2085	**0.1871**
AB5	**0.5937**	0.6007	**0.8547**	0.8750	**0.2515**	0.2548
BA1-5	0.7187	**0.6979**	1.1602	**1.0993**	0.3622	**0.3372**
BA2-5	0.7761	**0.7682**	1.4203	**1.4043**	0.4615	**0.4520**
CO1-5	0.4159	**0.4039**	0.4856	**0.4707**	0.1243	**0.1207**
CO2-5	0.3727	**0.3552**	0.4217	**0.3948**	0.1064	**0.0990**
HO10	0.7921	**0.7468**	1.6814	**1.4935**	0.2367	**0.2091**
AB10	0.7750	**0.7732**	**1.7603**	1.8014	**0.2531**	0.2571
BA1-10	0.8699	**0.8576**	2.4688	**2.3383**	0.3796	**0.3522**
BA2-10	0.8850	**0.8827**	2.8525	**2.8630**	0.4535	**0.4497**
CO1-10	0.6141	**0.6072**	1.0158	**0.9876**	0.1294	**0.1256**
CO2-10	0.5853	**0.5787**	0.8784	**0.8555**	0.1080	**0.1051**
Median difference	−0.0120		−0.02755		−0.01860	
p-value	0.00053		0.00329		0.00074	

or C-index to represent these relations. The C-index is computed as the proportion of the number of concordant pairs to the number of comparable pairs (see [15], page 50):

$$\text{C-index} = \frac{1}{\sum_{C_p \prec C_q} T_{C_p} T_{C_q}} \sum_{y_i \prec y_j} (\delta(y_i^* \prec y_j^*) + \frac{1}{2}\delta(y_i^* = y_j^*)), \tag{8}$$

where T_{C_p} and T_{C_q} are respectively the numbers of test examples with the class label C_p and C_q, $\{y_i, y_j\}$ is the real pair from the test examples, while $\{y_i^*, y_j^*\}$ is the corresponding predicted pair. When there are only two different class labels, the C-index amounts to the area under the Receiver Operating Characteristic (ROC) curve [16] and ranges from 0.5 to 1. A lower MZE or MAE means a better performance, while a higher C-index means a better performance. Here, we replace C-index by $(1 - \text{C-index})$ to keep an analogy with the other performance measures and facilitate the discussion of the results.

4.3 Performance Analysis

In this subsection, we analyze the performance of the proposed method on the different datasets listed in Subsect. 4.1. All the experimental results are obtained by applying ten-fold cross validation. We perform experiments on all datasets, setting the number k of nearest neighbor examples to 5. Table 2 shows the performance on *Dataset 1* and *Dataset 2*. It is clear that the performance on *Dataset 2* is better than the performance on *Dataset 1* for almost all original datasets except for *AB5* and *AB10*.

In order to test whether there is a significant difference in performance on these two datasets, we perform the Wilcoxon signed-rank test [12] at a significance level of $\alpha = 0.05$. If the p-value is smaller than the fixed significance level of α, then it means that there exists a statistically significant difference between these two datasets. In Table 2, it can be seen that the p-values for MZE, MAE and $1-$C-index are smaller than α, which means that there exists a statistically significant difference between the performance on these two datasets obtained from all original datasets. The experimental results, together with the obtained p-values and associated point estimates (median differences), show that using relative information with frequency distributions is meaningful.

5 Conclusions and Future Work

Based on our previous work [6], we have proposed an augmented method for ordinal classification for the setting in which there exists a small amount of absolute information and a large amount of relative information with frequency distributions. Specifically, we adapt the method of nearest neighbors for dealing with relative information with frequency distributions. We have carried out experiments on some classical ordinal classification or regression datasets. The experimental results show that the performance improves when relative information with frequency distributions is considered, which validates the usefulness of taking into account relative information with frequency distributions.

We see several interesting future directions for extending this work. On the one hand, absolute information with frequency distributions is also common. How to combine both absolute and relative information with frequency distributions for ordinal classification is still an open problem. On the other hand, in case the amount of relative information is large, it might be necessary to explore how to select the most informative pairwise comparisons for relative information in order to reduce the computational complexity of the proposed method.

Acknowledgements. Mengzi Tang is supported by the China Scholarship Council (CSC). Raúl Pérez-Fernández acknowledges the support of the Research Foundation of Flanders (FWO17/PDO/160) and the Spanish MINECO (TIN2017-87600-P). This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

References

1. Chen, J., Liu, X., Lyu, S.: Boosting with side information. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012. LNCS, vol. 7724, pp. 563–577. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37331-2_43
2. De Bie, T., Momma, M., Cristianini, N.: Efficiently learning the metric with side-information. In: Gavaldá, R., Jantke, K.P., Takimoto, E. (eds.) ALT 2003. LNCS (LNAI), vol. 2842, pp. 175–189. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-39624-6_15
3. Ovadia, S.: Ratings and rankings: reconsidering the structure of values and their measurement. Int. J. Soc. Res. Methodol. **7**(5), 403–414 (2004)
4. van Herk, H., van de Velden, M.: Insight into the relative merits of rating and ranking in a cross-national context using three-way correspondence analysis. Food Qual. Prefer. **18**(8), 1096–1105 (2007)
5. Sader, M., Verwaeren, J., Pérez-Fernández, R., De Baets, B.: Integrating expert and novice evaluations for augmenting ordinal regression models. Inf. Fusion **51**, 1–9 (2019)
6. Tang, M., Pérez-Fernández, R., De Baets, B.: Fusing absolute and relative information for augmenting the method of nearest neighbors for ordinal classification. Inf. Fusion **56**, 128–140 (2020)
7. Deza, M.M., Deza, E.: Encyclopedia of Distances, pp. 1–583. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-30958-8
8. Beliakov, G., Bustince, H., James, S., Calvo, T., Fernández, J.: Aggregation for Atanassov's intuitionistic and interval valued fuzzy sets: the median operator. IEEE Trans. Fuzzy Syst. **20**(3), 487–498 (2011)
9. Asuncion, A., Newman, D.: UCI Machine Learning Repository (2007). http://www.ics.uci.edu/~mlearn/MLRepository.html
10. PASCAL: Pattern Analysis, Statistical Modelling and Computational Learning. Machine Learning Benchmarks Repository (2011). http://mldata.org/
11. Chu, W., Ghahramani, Z.: Gaussian processes for ordinal regression. J. Mach. Learn. Res. **6**(7), 1019–1041 (2005)
12. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)
13. 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)
14. Baccianella, S., Esuli, A., Sebastiani, F.: Evaluation measures for ordinal regression. In: Proceedings of the 9th IEEE International Conference on Intelligent Systems Design and Applications, pp. 283–287. IEEE, Pisa (2009)
15. Waegeman, W., De Baets, B., Boullart, L.: Learning to rank: a ROC-based graph-theoretic approach. Pattern Recogn. Lett. **29**(1), 1–9 (2008)
16. Cortes, C., Mohri, M.: AUC optimization vs. error rate minimization. In: Proceedings of the 16th International Conference on Neural Information Processing Systems, pp. 313–320. MIT Press, Cambridge (2003)