Fisher Score-Based Feature Selection for Ordinal Classification: A Social Survey on Subjective Well-being[⋆]

M. Pérez-Ortiz[1], M. Torres-Jiménez[1], P.A. Gutiérrez[2], J. Sánchez-Monedero[1], and C. Hervás-Martínez[2]

[1] Universidad Loyola Andalucía, Dept. of Quantitative Methods, Córdoba, Spain
i82perom@uco.es,mtorres@uloyola.es,jsanchezm@uco.es
[2]University of Córdoba, Dept. of Computer Science and Numerical Analysis, Córdoba, Spain
{pagutierrez,chervas}@uco.es

Abstract. This paper approaches the problem of feature selection in the context of ordinal classification problems. To do so, an ordinal version of the Fisher score is proposed. We test this new strategy considering data from an European social survey concerning subjective well-being, in order to understand and identify the most important variables for a person's happiness, which is represented using ordered categories. The input variables have been chosen according to previous research, and these have been categorised in the following groups: demographics, daily activities, social well-being, health and habits, community well-being and personality/opinion. The proposed strategy shows promising results and performs significantly better than its nominal counterpart, therefore validating the need of developing specific ordinal feature selection methods. Furthermore, the results of this paper can shed some light on the human psyche by analysing the most and less frequently selected variables.

1 Introduction

The nature of well-being is a topic that has exercised the minds of moral philosophers for centuries [1]. Recently, research on happiness has gained importance, not only in the psychology area, but also in other fields like economics [2]. A number of nations have begun to develop measures of subjective well-being [3] to complement traditional measures of national well-being, such as the Gross Domestic Product. Well-being research is usually clustered into two camps [1], focusing either on subjective well-being or psychological well-being. On the one hand, subjective well-being is understood as having an emotional component of the balance between positive and negative affect and a cognitive component of judgements about life satisfaction. On the other hand, psychological well-being

[⋆] This work has been subsidised by the TIN2014-54583-C2-1-R project of the Spanish Ministerial Commission of Science and Technology (MICYT), FEDER funds and the P11-TIC-7508 project of the "Junta de Andalucía" (Spain).

has been defined as “engagement with existential challenges of life” [4]. Given the diversity of perspectives on the definition of subjective and psychological well-being, it is not surprising that different measurements have been considered in each case. In empirical research, a number of studies suggests that subjective and psychological well-being are two related, but distinct, constructs [4].

Nowadays, machine learning represents one of the most actively researched technical fields, mainly because of its applicability to very different domains. In this sense, machine learning, which lies at the intersection of computer science and statistics and is at the core of artificial intelligence and data science, addresses the question of how to build computer-based systems that improve automatically through experience. Given the lack of empirical agreement on the structure of well-being and the use of non-validated measures in previous studies, the current study examines these issues using machine learning techniques and data from the European Social Survey (ESS)[1]. The ESS includes a large sample of European inhabitants and validated well-being measures. It is an academically driven cross-national survey and has been conducted every two years across Europe. The survey measures the attitudes, beliefs and behaviour patterns of diverse populations in more than thirty nations, involving strict random probability sampling, a minimum target response rate of 70% and rigorous translation protocols. The interviews include questions on a variety of core topics: social trust, political interest and participation, socio-political orientations, social exclusion, national, ethnic and religious allegiances, health and social determinants, immigration, human values, demographics and socioeconomics.

This paper presents a new strategy to perform feature selection in ordinal classification [5]. Ordinal classification comprises those classification problems where the variable to predict follows a natural order (e.g. in Likert scales, as the case considered here). More specifically, we develop a novel feature selection methodology based on the well-known Fisher score [6]. The proposed technique promotes features that maintain the order among the classes and is used in this paper to analyse which are the factors influencing subjective well-being to a larger extent.

The remainder of the paper is structured as follows: Section II presents the data considered; Section III presents some previous notions and the proposed strategy for feature selection; Section IV analyses and presents the results obtained; and finally, Section V outlines some conclusions and final remarks.

2 Social survey on subjective well-being

The survey data conducted in 2014 from 15 European Union countries have been selected according to the availability of information (all persons aged 15 and over, resident within private households, regardless of their nationality, citizenship, language or legal status, in the following participating countries: Austria, Belgium, Czech Republic, Denmark, Estonia, Finland, France, Germany, Ireland, Netherlands, Slovenia, Sweden and Switzerland). Different variables have

[1] http://www.europeansocialsurvey.org/

been selected to predict the level of well-being of European citizens, considering the components that influences happiness according to previous research [7, 8]. These variables have been classified into six different groups:

- Demographics: different factors including country, age, gender, education, familiar composition, financial matters, etc.
- Daily activities: this group considers different variables that indicate how people spend their time (e.g. number of working hours, main activity, employment contract, number of hours watching TV, etc.).
- Social well-being: these variables are related to the social environment of the person (e.g. how often they take part in social activities, the number of people they are living with, how many people they can discuss personal matters with, etc.).
- Health and habits: including how often they practice sports, how often they eat vegetables, subjective general health, how often they drink alcohol, smoking behaviour, quality of sleep, and others.
- Community well-being: related to the sense of engagement they have with the area they live. It includes politic and environmental aspects (e.g. whether they feel close to their country, how interested they are in politics, how satisfied they are with economy/health services/education/government in their country, placement on left-right scale, etc.).
- Personality/opinion: how religious they are, whether it is important to be rich/free/humble/adventurous, whether they would allow immigrants from poorer countries to settle in their country, whether most people can be trusted or not, etc.

The study comprise a set of 56 different variables: a large number of them (38) represent Likert scales (i.e. ordinal) and are codified using a numeric scale, 7 are numeric, 6 of them are binary, and finally, there are 5 nominal variables, which are transformed to binary ones, resulting then in a total set of 91 variables. The total number of cases is $28,137$, excluding those with responses "don't know", "no answer" or "refusal" in the dependent variable, which is how happy they are in a Likert scale (from 0 to 10). Three different datasets are considered, using different number of levels for the subjective well-being: all the 11 possible answers (referred to as SW-11c), 5 classes (SW-5c, where $\mathcal{C}_1 = \{0, 1, 2\}$, $\mathcal{C}_2 = \{3, 4\}$, $\mathcal{C}_3 = \{5, 6\}$, $\mathcal{C}_4 = \{7, 8\}$ and $\mathcal{C}_5 = \{9, 10\}$) and 3 classes (SW-3c, where $\mathcal{C}_1 = \{0, 1, 2, 3\}$, $\mathcal{C}_2 = \{4, 5, 6, 7\}$ and $\mathcal{C}_3 = \{8, 9, 10\}$).

Values such as "don't know", "no answer" or "refusal" in the independent variables have been considered as missing values and have been imputed using the Event Covering algorithm [9], as suggested in [10] for approximate models such as neural networks, support vector machines and other statistical methods.

3 Methodology

This section describes some previous notions (the paradigm of ordinal classification and the Fisher score) and proposes an ordinal feature selection method.

3.1 Previous notions

Ordinal classification and ordinal feature selection The classification of patterns into naturally ordered labels is referred to as ordinal regression or ordinal classification. This learning paradigm, although still mostly unexplored, is spreading rapidly and receiving a lot of attention from the pattern recognition and machine learning communities [5, 11], given its applicability to real world problems. This paradigm shares properties of classification and regression. In contrast to multinomial classification, there exists some ordering among the elements of $\mathcal{Y}$ (the labelling space) and both standard classifiers and the zero-one loss function do not capture and reflect this ordering. Concerning regression, $\mathcal{Y}$ is a non-metric space, so the distances between categories are not known.

As an explanatory example, consider the case of financial trading where an agent predicts not only whether to buy an asset, but also the investment. The different situations could be categorised as {"no investment", "little investment", "big investment", "huge investment"}. A natural order among the classes exists in this case, and a necessity of penalising differently the misclassification errors (it should not be considered equal misclassifying a "no investment" instance with a "huge investment" one than misclassifying it with "little investment").

The goal in ordinal classification is to assign an input vector $\mathbf{x}$ to one of K discrete classes $\mathcal{C}_k, k \in \{1, \ldots, K\}$, where there exists a given ordering between the labels, $\mathcal{C}_1 \prec \mathcal{C}_2 \prec \cdots \prec \mathcal{C}_K$, $\prec$ denoting this order information. Hence, the objective is to find a prediction rule $C : \mathcal{X} \rightarrow \mathcal{Y}$ by using an i.i.d. training sample $X = \{\mathbf{x}_i, y_i\}_{i=1}^N$, where N is the number of training patterns, $\mathbf{x}_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, $\mathcal{X} \subset \mathbb{R}^d$ is the d-dimensional input space and $\mathcal{Y} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K\}$ is the label space. For convenience, denote by $\mathbf{X}_i$ to the set of patterns belonging to $\mathcal{C}_i$.

Despite the novelty of ordinal classification, there is some research concerning new prediction strategies for these problems. However, there are aspects of ordinal classification that are receiving less attention. This is the case of feature selection methods in ordinal classification, where the number of approaches is still low [12, 13]. Like some of the strategies in the feature selection literature, these techniques rely on a discretisation of the input space. In this paper, we devise a new strategy for ordinal feature selection that is free of this requirement.

Fisher score for feature selection The problem of supervised feature selection is now described. Given a dataset $\{\mathbf{x}_i, y_i\}_{i=1}^N$, we aim to find a feature subset of size m (where $m < d$) that contains the most informative features.

The Fisher score for feature selection [6] was proposed as an heuristic strategy for computing an independent score for each feature using the well-known notion of the Fisher ratio. Let μ_k^i and σ_k^i be the mean and standard deviation of the k-th class and i-th feature (and μ^i and σ^i the mean and standard deviation of the whole dataset for the i-th feature). The Fisher score for the i-th feature ($\mathbf{x}^i$) can be computed as:

$$F(\mathbf{x}^i) = \frac{\sum_{k=1}^K N_k(\mu_k^i - \mu^i)^2}{\sum_{k=1}^K N_k(\sigma_k^i)^2}, \tag{1}$$

where N_k is the number of patterns of class $\mathcal{C}_k$. Since this score is computed independently, the features selected may represent a suboptimal set. Furthermore, this heuristic may fail to select redundant features or those with a high aggregated discriminative power. This technique is named as Nominal Feature Selection (NFS) in the experimental results.

3.2 Proposed feature selection strategy

In this paper, we reformulate the nominal Fisher score F to deal with ordinal data (named it as F_{O}). In this regard, we include a weighting term that introduces a higher cost for distant classes. This cost will force the feature selection method to focus more on those features that help to discriminate classes that are far in the ordinal scale (in order to avoid the above-mentioned misclassification errors). The formulation proposed is the maximise the following score:

$$F_{\mathrm{O}}(\mathbf{x}^i) = \frac{\sum_{k=1}^{K}\sum_{j=1}^{K}|k-j|(\mu_k^i - \mu_j^i)^2}{(K-1)\sum_{k=1}^{K}(\sigma_k^i)^2}, \tag{2}$$

where $|k-j|$ is the cost of misclassifying a pattern from class $\mathcal{C}_k$ in class $\mathcal{C}_j$.

Apart from the fact that more distant classes in the ordinal scale should present a higher distance between them, there is another ordinal requirement which can be introduced in the feature selection stage. As said before, the labelling space is non-metric, therefore we can not introduce a distance relation among the different classes. However, from the ordinal classification definition it can be stated that $d(\mathcal{C}_k,\mathcal{C}_j) < d(\mathcal{C}_k,\mathcal{C}_h)$, $\forall\{k,j,h,|k\neq j\neq h \wedge (k<j<h \vee k>j>h)\}$. These ordinal requirements can be introduced by the score:

$$O_{\mathrm{R}}(\mathbf{x}^i) = \frac{\sum_{k=1}^{K-2}\sum_{j=k+1}^{K-1}\sum_{h=j+1}^{K}[\![((\mu_k^i - \mu_h^i)^2 - (\mu_k^i - \mu_j^i)^2) > 0]\!]}{\sum_{j=2}^{K-1}(K-j)}, \tag{3}$$

being $[\![\cdot]\!]$ a Boolean test which is 1 if the inner condition is true, and 0 otherwise. This O_{R} score measures the number of ordinal requirements fulfilled for a specific feature i. To include both terms in the feature selection process, we compute a weighted mean of both scores in the following manner:

$$F_{\mathrm{OR}}(\mathbf{x}^i) = \alpha \cdot F_{\mathrm{O}}(\mathbf{x}^i) + (1-\alpha)O_{\mathrm{R}}(\mathbf{x}^i), \tag{4}$$

where $\alpha \in (0,1)$.

Up to now, we have defined the distance between the classes $\mathcal{C}_z$ and $\mathcal{C}_j$ as $d_{\mathrm{s}}^i(\mathcal{C}_z,\mathcal{C}_j) = (\mu_z^i - \mu_j^i)^2$ (note that this formulation presents problems with non-linear, multimodal or non-normal data). Alternative metrics have been proposed for measuring the distance between classes [14]. In this way, we consider other notions of distance between sets of points, such as the mean distance:

$$d_{\mathrm{m}}^i(\mathcal{C}_z,\mathcal{C}_j) = \frac{1}{N_z \cdot N_j}\sum_{\mathbf{x}_h^i\in\mathcal{C}_z}\sum_{\mathbf{x}_v^i\in\mathcal{C}_j}(x_h^i - x_v^i)^2, \tag{5}$$

where the idea is to compute the mean distance between each pair of patterns of different classes. Another alternative is the sum of minimum distances:

$$d^i_{\mathrm{md}}(\mathcal{C}_z, \mathcal{C}_j) = \frac{1}{2} \left(\sum_{\mathbf{x}^i_h \in \mathcal{C}_z} \mathit{\Delta}_{\mathrm{m}}(\mathbf{x}^i_h, \mathcal{C}_j) + \sum_{\mathbf{x}^i_v \in \mathcal{C}_j} \mathit{\Delta}_{\mathrm{m}}(\mathbf{x}^i_v, \mathcal{C}_z) \right), \tag{6}$$

where $\mathit{\Delta}_{\mathrm{m}}(\mathbf{x}^i_h, \mathcal{C}_j) = \min_{\mathbf{x}^i_v \in \mathcal{C}_j} d(\mathbf{x}^i_h, \mathbf{x}^i_v)$ and d represents the Euclidean distance. Finally, the Hausdorff distance is defined as:

$$d^i_{\mathrm{h}}(\mathcal{C}_z, \mathcal{C}_j) = \max\{ \max_{\mathbf{x}^i_h \in \mathcal{C}_z} \mathit{\Delta}_{\mathrm{m}}(\mathbf{x}^i_h, \mathcal{C}_j), \max_{\mathbf{x}^i_v \in \mathcal{C}_j} \mathit{\Delta}_{\mathrm{m}}(\mathbf{x}^i_v, \mathcal{C}_z) \}. \tag{7}$$

In the experiments of this paper, we will test these three alternative approaches for computing inter-class distances.

4 Experimental results

This section exposes the experiments considered in this paper and analyses the results obtained. Regarding the experimental setup, a stratified holdout technique was applied to randomly divide the datasets 30 times, using 5% of the patterns for training (approximately 1400 patterns) and the remaining 95% for testing. The partitions were the same for all methods and one model was obtained and evaluated (in the test set) for each split. Finally, the results are computed as the mean and standard deviation of the measures over the 30 test sets. The classification method chosen is the reformulation of Kernel Discriminant Analysis for Ordinal Regression (KDLOR), because of its relation to the Fisher score [15].

4.1 Methodologies tested

Different methods are compared:

- No feature selection. These results are obtained with the KDLOR classification method using the whole set of features.
- Nominal Feature Selection (NFS). In this case, we use the nominal version of the Fisher score.
- Ordinal Feature Selection (OFS) with d_{s} as distance metric.
- Ordinal Feature Selection using the mean distance (OFSMean). In this case, d_{m} is used as distance metric.
- Ordinal Feature Selection using the min distance (OFSMin). In this case, d_{md} is used as distance metric.
- Ordinal Feature Selection using the Hausdorff distance (OFSHausdorff). In this case, d_{h} is used as distance metric.

The value α for all the ordinal feature selection methods is fixed to $\alpha = 0.5$. For all feature selection methods, features are ranked according to the corresponding score, and then a percentage of the best features is retained (where the percentages tested are $10\%, 20\%, \ldots, 90\%$). Concerning the parameters selected for the classification method, a Gaussian kernel is used for KDLOR, $K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma^2}\right)$, where σ is the kernel width which has been cross-validated using a 5-fold nested procedure over the training set and the range $\{10^{-3}, \ldots, 10^{3}\}$.

4.2 Evaluation metrics

Several measures can be considered for evaluating ordinal classifiers, such as the Mean Absolute Error (MAE) or the accuracy or Acc [5]. In this work, two metrics have been used:

- Acc is the ratio of correctly classified patterns:

$$Acc = \frac{1}{N} \sum_{i=1}^{N} [\![y_i^* = y_i]\!],$$

  where y_i is the desired output for pattern i and y_i^* is the prediction.
- Acc and MAE may not be the best option when measuring performance in the presence of class imbalances [16]. The average mean absolute error ($AMAE$) 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. Let MAE_k be the MAE for a given k-th class:

$$MAE_k = \frac{1}{N_k} \sum_{i=1}^{N_k} |\mathcal{O}(y_i) - \mathcal{O}(y_i^*)|,\ 1 \leq k \leq K,$$

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

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

  MAE values range from 0 to $K-1$, as do those of $AMAE$. This metric has been chosen given the imbalanced nature of the problem considered.

4.3 Results

The results obtained for all the methods can be seen in Table 1 for Acc and Table 2 for $AMAE$, from which several conclusions can be drawn. Note that the results without feature selection are also included in the Tables for the three datasets. Firstly, the performance of the base algorithm (i.e. with no feature selection) can be improved in some cases (e.g. in SW-5c), and there are reduction levels where the reduced datasets perform relatively similar to the base ones (e.g. with 30% of features), which is interesting for model interpretability purposes and to reduce

Table 1. *Acc* mean and standard deviation (Mean ± SD) obtained by all the methodologies compared as a function of the percentage of features selected.

	NFS	OFS	OFSMean	OFSMin	OFSHausdorff
Perc. of features	SW-3c, result without FS: 68.99 ± 0.53				
10%	46.31 ± 2.55	49.02 ± 3.75	*49.41 ± 2.29*	45.11 ± 4.67	**55.52 ± 4.55**
20%	**56.83 ± 4.45**	*56.76 ± 6.28*	53.14 ± 1.41	49.59 ± 2.46	49.99 ± 2.58
30%	60.25 ± 0.21	*60.35 ± 0.09*	52.36 ± 1.32	60.14 ± 0.57	**65.26 ± 0.67**
40%	*60.34 ± 0.27*	60.33 ± 0.18	56.69 ± 1.31	62.14 ± 0.83	**65.01 ± 0.80**
50%	60.40 ± 0.07	*60.41 ± 0.05*	60.32 ± 0.12	64.21 ± 0.71	**64.22 ± 0.65**
60%	59.93 ± 1.00	60.05 ± 0.82	59.53 ± 1.01	*66.48 ± 0.66*	**66.49 ± 0.63**
70%	60.36 ± 0.51	60.77 ± 0.74	61.74 ± 0.69	**67.37 ± 0.62**	*67.36 ± 0.68*
80%	62.91 ± 0.52	63.46 ± 0.68	64.27 ± 0.69	**68.39 ± 0.51**	*68.28 ± 0.48*
90%	66.09 ± 0.52	66.18 ± 0.51	66.98 ± 0.64	*68.87 ± 0.50*	**68.89 ± 0.54**
Perc. of features	SW-5c, result without FS: 50.22 ± 0.49				
10%	*30.38 ± 3.49*	**32.41 ± 4.24**	26.62 ± 9.84	22.10 ± 3.26	22.73 ± 3.16
20%	38.15 ± 10.35	**44.75 ± 7.18**	29.63 ± 3.69	*42.32 ± 3.73*	42.28 ± 3.73
30%	*48.67 ± 0.25*	48.40 ± 2.24	38.17 ± 1.67	**48.84 ± 0.06**	**48.84 ± 0.06**
40%	48.82 ± 0.07	*48.84 ± 0.04*	42.84 ± 0.83	**48.86 ± 0.02**	**48.86 ± 0.02**
50%	*48.84 ± 0.01*	**48.85 ± 0.01**	47.65 ± 0.50	48.69 ± 0.42	48.67 ± 0.43
60%	48.85 ± 0.00	48.85 ± 0.00	48.85 ± 0.01	*49.72 ± 0.59*	**50.74 ± 0.66**
70%	48.82 ± 0.18	48.84 ± 0.05	48.78 ± 0.18	*50.80 ± 0.41*	**51.10 ± 0.40**
80%	48.88 ± 0.22	49.05 ± 0.35	49.51 ± 0.34	*50.88 ± 0.31*	**51.48 ± 0.22**
90%	49.88 ± 0.33	49.94 ± 0.30	50.21 ± 0.27	*50.42 ± 0.23*	**51.08 ± 0.20**
Perc. of features	SW-11c, result without FS: 30.09 ± 0.34				
10%	10.85 ± 3.72	**15.82 ± 3.27**	6.70 ± 4.80	14.35 ± 2.34	*14.60 ± 2.56*
20%	17.97 ± 6.07	*24.58 ± 6.61*	13.06 ± 2.86	22.52 ± 3.10	**27.88 ± 4.08**
30%	*30.49 ± 0.30*	29.80 ± 3.12	19.30 ± 1.07	**30.71 ± 0.08**	30.33 ± 1.14
40%	*30.59 ± 0.41*	**30.60 ± 0.49**	21.83 ± 1.46	29.07 ± 2.37	27.78 ± 0.47
50%	*30.51 ± 0.28*	**30.61 ± 0.20**	30.01 ± 1.35	27.54 ± 0.86	29.38 ± 0.41
60%	**30.73 ± 0.03**	**30.73 ± 0.04**	*30.35 ± 1.32*	28.77 ± 0.57	30.18 ± 0.36
70%	*29.47 ± 1.99*	28.51 ± 2.01	27.39 ± 1.26	29.41 ± 0.47	**30.37 ± 0.31**
80%	27.66 ± 0.37	28.13 ± 0.48	28.53 ± 0.51	*29.79 ± 0.34*	**30.20 ± 0.35**
90%	29.01 ± 0.28	29.21 ± 0.27	29.44 ± 0.40	*29.98 ± 0.35*	**30.11 ± 0.31**
Average	43.777	44.639	41.233	*44.977*	**45.838**
Ranking	3.519	2.778	3.981	*2.759*	**1.963**
Friedman's test	Confidence interval $C_0 = (0, F_{(\alpha=0.05)} = 2.459)$. F-val.$_{Acc}$: $8.278 \notin C_0$.				

The best performing method is in **bold** face and the second one in *italics*.

computational time. Secondly, the results are in general promising (considering the difficulty of the problem). Finally, the proposed technique performs better than the nominal counterpart (specially when the Hausdorff distance is used).

To quantify whether a statistical difference exists among the algorithms compared, a procedure is employed to compare multiple classifiers in multiple datasets [17]. Tables 1 and 2 also shows the result of applying the non-parametric statistical Friedman's test (for a significance level of $\alpha = 0.05$) to the mean *Acc* and *AMAE* rankings. The test rejected the null-hypothesis that all of the algorithms perform similarly in mean ranking for both metrics.

On the basis of this rejection and following the guidelines in [17], we consider the best performing method as control method for the following test. We compare this method to the rest according to their rankings. It has been noted that the approach of comparing all classifiers to each other in a post-hoc test is not as sensitive as the approach of comparing all classifiers to a given classifier (a control method). One approach to this latter type of comparison is the Holm's test. The test statistics for comparing the i-th and j-th method using this procedure is:

Table 2. *AMAE* mean and standard deviation (Mean ± SD) obtained by all the methodologies compared as a function of the percentage of features selected.

	NFS	OFS	OFSMean	OFSMin	OFSHausdorff
Perc. of features	SW-3c, result without FS: 0.581 ± 0.010				
10%	0.877 ± 0.037	0.891 ± 0.039	0.870 ± 0.036	*0.825 ± 0.049*	**0.780 ± 0.054**
20%	0.947 ± 0.057	0.955 ± 0.084	**0.756 ± 0.020**	*0.757 ± 0.035*	0.758 ± 0.052
30%	0.996 ± 0.006	0.999 ± 0.002	*0.731 ± 0.019*	0.917 ± 0.105	**0.606 ± 0.108**
40%	0.997 ± 0.009	0.997 ± 0.007	0.913 ± 0.046	*0.647 ± 0.014*	**0.606 ± 0.014**
50%	0.997 ± 0.003	0.999 ± 0.002	0.997 ± 0.006	*0.631 ± 0.009*	**0.630 ± 0.009**
60%	0.952 ± 0.098	0.950 ± 0.102	*0.806 ± 0.140*	**0.610 ± 0.007**	**0.610 ± 0.008**
70%	0.777 ± 0.114	0.730 ± 0.075	*0.677 ± 0.020*	**0.600 ± 0.007**	**0.600 ± 0.008**
80%	0.678 ± 0.016	0.668 ± 0.016	*0.645 ± 0.017*	**0.588 ± 0.009**	*0.589 ± 0.009*
90%	0.630 ± 0.014	0.629 ± 0.015	*0.609 ± 0.014*	**0.581 ± 0.008**	**0.581 ± 0.008**
Perc. of features	SW-5c, result without FS: 1.294 ± 0.052				
10%	*1.619 ± 0.098*	**1.590 ± 0.107**	1.668 ± 0.192	1.678 ± 0.116	1.668 ± 0.097
20%	1.490 ± 0.199	1.460 ± 0.168	1.486 ± 0.085	*1.349 ± 0.080*	**1.347 ± 0.079**
30%	*1.401 ± 0.001*	1.413 ± 0.070	1.404 ± 0.023	**1.399 ± 0.001**	**1.399 ± 0.001**
40%	**1.400 ± 0.001**	**1.400 ± 0.000**	*1.410 ± 0.014*	**1.400 ± 0.000**	**1.400 ± 0.000**
50%	1.400 ± 0.000	1.400 ± 0.000	1.401 ± 0.005	*1.356 ± 0.100*	**1.339 ± 0.112**
60%	1.400 ± 0.000	1.400 ± 0.000	1.400 ± 0.000	*1.160 ± 0.049*	**1.137 ± 0.066**
70%	1.396 ± 0.022	1.396 ± 0.023	1.370 ± 0.062	*1.192 ± 0.021*	**1.145 ± 0.023**
80%	1.358 ± 0.050	1.336 ± 0.054	1.280 ± 0.045	*1.249 ± 0.017*	**1.181 ± 0.017**
90%	1.304 ± 0.028	1.297 ± 0.022	*1.284 ± 0.016*	1.294 ± 0.010	**1.232 ± 0.010**
Perc. of features	SW-11c, result without FS: 2.884 ± 0.015				
10%	3.986 ± 0.353	**3.774 ± 0.255**	4.530 ± 0.634	*3.381 ± 0.229*	3.384 ± 0.200
20%	3.670 ± 0.425	3.549 ± 0.146	3.696 ± 0.333	**3.143 ± 0.145**	*3.404 ± 0.215*
30%	3.535 ± 0.011	3.514 ± 0.111	**3.298 ± 0.150**	3.546 ± 0.001	*3.494 ± 0.143*
40%	3.535 ± 0.025	3.537 ± 0.030	*3.246 ± 0.087*	3.264 ± 0.403	**2.656 ± 0.049**
50%	3.528 ± 0.018	3.534 ± 0.017	3.520 ± 0.076	*2.767 ± 0.157*	**2.685 ± 0.040**
60%	3.543 ± 0.003	3.542 ± 0.004	3.497 ± 0.167	*2.765 ± 0.047*	**2.739 ± 0.030**
70%	3.386 ± 0.248	3.222 ± 0.289	2.943 ± 0.212	*2.802 ± 0.035*	**2.799 ± 0.025**
80%	2.976 ± 0.024	2.929 ± 0.036	2.861 ± 0.046	**2.843 ± 0.021**	*2.848 ± 0.020*
90%	2.913 ± 0.022	2.905 ± 0.023	2.875 ± 0.028	*2.870 ± 0.018*	**2.867 ± 0.016**
Average	1.914	1.889	1.858	*1.712*	**1.647**
Ranking	4.185	3.852	3.093	*2.370*	**1.500**
Friedman's test	Confidence interval $C_0 = (0, F_{(\alpha=0.05)} = 2.459)$. F-val.$_{AMAE}$: $23.859 \notin C_0$.				

The best performing method is in **bold** face and the second one in *italics*.

$z = \frac{R_i - R_j}{\sqrt{\frac{J(J+1)}{6T}}}$, where J is the number of algorithms, T is the number of datasets and R_i is the mean ranking of the i-th method. The z value is used to find the corresponding probability from the table of the normal distribution, which is then compared with an appropriate level of significance α. Holm's test adjusts the value for α in order to compensate for multiple comparisons. This is done in a step-up procedure that sequentially tests the hypotheses ordered by their significance. We will denote the ordered p-values by $p_1, p_2, \ldots, p_q$ so that $p_1 \leq p_2 \leq \ldots \leq p_q$. Holm's test compares each p_i with $\alpha^*_{\text{Holm}} = \alpha/(J-i)$, starting from the most significant p value. If p_1 is below $\alpha/(J-1)$, the corresponding hypothesis is rejected, and we allow to compare p_2 with $\alpha/(J-2)$. If the second hypothesis is rejected, the test proceeds with the third, and so on.

Table 3 shows the result of the Holm's test when comparing the best performing technique (i.e. OFSHausdorff) to the rest of algorithms. It can be seen that this method outperforms the rest of techniques for *AMAE* and specifically OFSMean and NFS for *Acc*, when $\alpha = 0.05$. This result validates the Hausdorff distance in this context and shows that the ordinal nature of the data should be considered when performing feature selection. The results not only improve when considering ordinal measures (i.e. $AMAE$) but also nominal ones (such as Acc), meaning that the method can benefit from the order constraint introduced.

Table 3. Results of the Holm procedure using OFSHausdorff as control method: corrected α values, compared method and p-values, ordered by comparisons (i).

Control alg.: OFSHausdorff		Acc	
i	$\alpha^*_{0.05}$	Method	p_i
1	0.01250	OFSMean	0.00000_{++}
2	0.01667	NFS	0.00003_{++}
3	0.02500	OFS	0.05830
4	0.05000	OFSMin	0.06425
Control alg.: OFSHausdorff		$AMAE$	
i	$\alpha^*_{0.05}$	Method	p_i
1	0.01250	NFS	0.00000_{++}
2	0.01667	OFS	0.00000_{++}
3	0.02500	OFSMean	0.00002_{++}
4	0.05000	OFSMin	0.04312_{++}

Win (++) with statistical significant difference for $\alpha = 0.05$

4.4 Discussion

As stated before, the authors consider that a percentage of 30% features represents a good option, since it allows to have a relatively accurate model with an increased interpretability (70% of the variables are removed). In this section, we examine the variables (associated to this 30%) that are of largest importance to the characterisation of happiness, considering the version of the dataset with 3 classes and the OFSHausdorff method. Note that the feature selection method does not consider aggregated sets of features, and therefore is not able to discover redundant variables of interactions between them. Since we considered 30 different train data partitions, we have 30 different results. The following variables have been selected at least in 15 models: variables related to the country (meaning that there could be other factors of vital importance to the characterisation of well-being, such as the state of economy of the country, the weather, etc.), gender, variables related to love relationships (single, legally married, living with partner, etc.), general level of health, whether the person is daily hampered by illness or if he/she belongs to a minority ethnic, the number of times the person felt depressed previous week, the quality of the sleep, whether the person gives importance to be rich and to have expensive belongings, religiosity and whether the person would allow entry of immigrants from poorer countries.

Several conclusions can be drawn from these results. Firstly, the environment in which a person lives significantly influences their well-being and so do other demographic variables (e.g. familiar composition, gender or belonging to a minority ethnic). Secondly, both physical and mental health have an impact on subjective well-being. Finally, the group of variables related to personality and opinions also play a vital role in the classification of happiness.

The variables selected when considering 70% of features were also examined in order to study variables that were selected by very few models. Note from Table 1 that 70% represents also an interesting threshold, as it is the point from which the performance usually starts to degrade. In this case, the range of variables present in at least in 15 models includes additional variables concerning: social well-being (feeling of loneliness and inability to get going, number of people to discuss personal matters with, involvement in social activities and conflicts at home when growing up), health and habits (diet/sport/alcohol/smoking behaviour, inability to get medical consultation or treatment), daily activity (specifically the main activity, the number of hours watching TV per week and whether the person improved their knowledge/skills in the last year), other demographic factors (such as the domicile, the type of contract or the familiar composition), community well-being (interest in politics, closeness to country, satisfaction with health and education in country) and, finally, personality (whether it is important to be modest and to seek for adventures, whether the person thinks that other people try to take advantage of them and the feeling of safety walking alone in local area at night). Conversely, there is a set of variables that have been selected in very few models (in this case we consider variables selected in less than 10 models). These variables are the following: age, weight, number of hours working, people responsible for at job, financial difficulties when growing up, years of education, number of people living with, satisfaction with government and economy (as opposed to health and education that were selected as relevant), whether politicians care what people think and the number of hours helping others (family, friends or neighbours).

5 Conclusions and future work

This paper presents a feature selection strategy for classification problems where the dependent variable follows a natural order. We construct a dataset for predicting subjective well-being across different European countries that includes 56 variables of different components of happiness. The results show that there are some factors, such as the environment where the person lives, the physical and mental health and the personality, that are of great influence to subjective well-being. Moreover, the performance of the proposed method is competitive against its nominal counterpart, which demonstrates the necessity of developing more specific techniques for domains such as the ordinal classification one.

Future research comprises including other countries (e.g. developing ones), performing a sensitivity analysis, comparing the features selected for the different versions of the dataset and extending the proposed methodology to consider aggregated sets of features.

References

1. Linley, P.A., Maltby, J., Wood, A.M., Osborne, G., Hurling, R.: Measuring happiness: The higher order factor structure of subjective and psychological well-being measures. Personality and Individual Differences **47**(8) (2009) 878 – 884

2. Diener, E.: Subjective well-being: The science of happiness and a proposal for a national index. American Psychologist **55**(1) (2000) 34–43
3. Self, A., Thomas, J., Randall, C.: Measuring national well-being: Life in the uk (2012) Last access: 8 dec 2015.
4. Keyes, C.L., Shmotkin, D., Ryff, C.D.: Optimizing well-being: the empirical encounter of two traditions. Journal of personality and social psychology **82**(6) (2002) 1007
5. Gutiérrez, P.A., Pérez-Ortiz, M., Sánchez-Monedero, J., Fernández-Navarro, F., Hervás-Martínez, C.: Ordinal regression methods: Survey and experimental study. Knowledge and Data Engineering, IEEE Transactions on **28**(1) (Jan 2016) 127–146
6. Gu, Q., Li, Z., Han, J.: Generalized fisher score for feature selection. CoRR **abs/1202.3725** (2012)
7. Bixter, M.T.: Happiness, political orientation, and religiosity. Personality and Individual Differences **72** (2015) 7 – 11
8. Hills, P., Argyle, M.: The Oxford Happiness Questionnaire: a compact scale for the measurement of psychological well-being. Personality and Individual Differences **33**(7) (November 2002) 1073–1082
9. Chiu, D., Wong, A.: Synthesizing knowledge: A cluster analysis approach using event-covering. IEEE Transactions on Systems and Man and Cybernetics and Part B **16**(2) (1986) 251–259
10. Luengo, J., García, S., Herrera, F.: On the choice of the best imputation methods for missing values considering three groups of classification methods. Knowledge and Information Systems **32**(1) (2012) 77–108
11. Pérez-Ortiz, M., Gutiérrez, P.A., Hervás-Martínez, C.: Projection-based ensemble learning for ordinal regression. Cybernetics, IEEE Transactions on **44**(5) (2014) 681–694
12. Baccianella, S., Esuli, A., Sebastiani, F.: Feature selection for ordinal text classification. Neural Comput. **26**(3) (March 2014) 557–591
13. Mukras, R., Wiratunga, N., Lothian, R., Chakraborti, S., Harper, D.: Information gain feature selection for ordinal text classification using probability redistribution. In: the IJCAI'07 Workshop on Text Mining and Link Analysis, Hyderabad, IN (2007)
14. Eiter, T., Mannila, H.: Distance measures for point sets and their computation. Acta Informatica **34** (1997) 103–133
15. Sun, B.Y., Li, J., Wu, D.D., Zhang, X.M., Li, W.B.: Kernel discriminant learning for ordinal regression. IEEE Transactions on Knowledge and Data Engineering **22** (2010) 906–910
16. 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 09), Pisa, Italy (December, 2009 2009)
17. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research **7** (2006) 1–30