INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

How Ordinal Are Your Data?

Sadari Jayawardena[1], *Julien Epps*[1,2], *Zhaocheng Huang*[1]

[1] School of Electrical Engineering and Telecommunications, UNSW Sydney, Australia
[2]Data61, CSIRO, Australia
s.jayawardena@unsw.edu.au, j.epps@unsw.edu.au

Abstract

Many affective computing datasets are annotated using ordinal scales, as are many other forms of ground truth involving subjectivity, e.g. depression severity. When investigating these datasets, the speech processing community has chosen classification problems in some cases, and regression in others, while ordinal regression may also arguably be the correct approach for some. However, there is currently essentially no guidance on selecting a suitable machine learning and evaluation method. To investigate this problem, this paper proposes a neural network-based framework which can transition between different modelling methods with the help of a novel multi-term loss function. Experiments on synthetic datasets show that the proposed framework is empirically well-behaved and able to correctly identify classification-like, ordinal regression-like and regression-like properties within multidimensional datasets. Application of the proposed framework to six real datasets widely used in affective computing and related fields suggests that more focus should be placed on ordinal regression instead of classifying or predicting, which are the common practices to date.

Index Terms: Ordinal Regression, Affective Computing, Data Ordinality

1. Introduction

The type of algorithms adopted for a given problem (i.e. classification, regression, etc.) normally depends on the nature of the ground truth. Because of their subjectivity, the ground truth for psychological and physiological research into depression prediction, emotion recognition and cognitive load monitoring is often derived from annotations on ordinal scales. Despite the ground truth being ordinal, in speech processing literature, such datasets are treated either as a classification or a regression problem [1-5], or both [6], presumably due to the well-established classification/regression frameworks readily accessible to the community. Nevertheless, some of these problems should arguably be tackled using ordinal regression, which is a kind of compromise between classification and regression. However, there is no obvious guideline with which to make a decision on appropriate machine learning and evaluation method.

This exploratory study aims to provide guidance on how to select the learning approach. Specifically, we propose a neural network-based framework that trades off between classification, ordinal regression and regression loss functions, with a particular focus on speech-based affective computing problems, which have been solved in multiple different ways to date.

2. Relation to Prior Work

The ground truth of a large number of psychological and physiological datasets are essentially ordinal. For example, depression scores such as Patient Health Questionnaire (PHQ-8) [7], Beck Depression Inventory (BDI-II) [8], adopted as ground truth in depression corpora, are determined using tools which consist of a Likert-scale questionnaire on symptoms of depression. Scores for sub-questions/symptoms are summed to produce the final depression score which implies a relative ordering and does not provide any numerical interpretation. Emotion dimensions (e.g. arousal, valence, dominance, etc) are usually annotated manually, on either a discrete [9, 10] or continuous basis [11]. Such annotation is highly sensitive to personal subjectivity, and hence it was argued in [12, 13] that emotion attributes should be considered ordinal rather than numerical. Similarly, the level of cognitive load is determined manually based on a subjective measure indicating the relative ordering of mental load [1, 14], and hence is ordinal. Despite the aforementioned ordinality in the ground truth labels, relatively less work has been done to investigate/exploit the ordinality in these areas.

The naïve, but widely adopted, method to approaching a problem that has been defined on an ordinal scale is to assume that either classification or regression will suffice, depending perhaps on the number of classes. When fewer classes are involved, such as for depression severity (e.g. *normal*, *mild*, *moderate*, *moderately-severe*, *severe*), discrete emotion categories (e.g. *anger*, *happiness*, *sadness*, *fear*, *surprise*) or cognitive load levels (e.g. *low*, *medium*, *high*), classification is often employed [1-3], whereas when the scale size is large or continuous, regression models are often applied [4, 5]. In this approach, ordinal scales (*discrete* and *ordered*) are implicitly converted into nominal (*discrete* and *unordered*) or numerical (*continuous* and *ordered*) representations respectively. Such transformations, however, could lose important information (i.e. classification ignores ordering between classes) or impose unrealistic assumptions (i.e. regression assumes a metric interpretation on class labels) [15].

Ordinal regression is considered as intermediate between classification and regression because labels are discrete and ordered [16, 17]. Recently, a few studies have taken the initiative to investigate ordinal regression, which is the fundamentally correct approach for ordinal problems [15, 18]. An empirical comparison of the three machine learning approaches to understand the practical advantage of ordinal regression has been undertaken in [19, 20]. Most ordinal regression models, including the proportional-odds model (POM) [21] and RankSVM [19] are not capable of handling large, high-dimensional datasets and hence their usage is limited to statistical and medical research where the datasets are small, compared with affective computing. An exception is a

http://dx.doi.org/10.21437/Interspeech.2020-2030

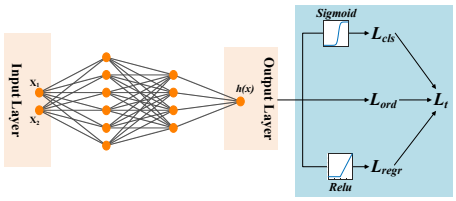

Figure 1: *Proposed neural network architecture that can gradually move between classification, ordinal regression and regression. The loss layer is in blue.*

preference neural network herein referred to as PrefNet [22], which can overcome the aforementioned problem with online loss calculation. A few benefits of neural network approaches in general include not imposing any restrictions on input data, non-linear data processing, generalizability and fault tolerance, which have made it well-recognised in many machine learning applications including ordinal neural networks [22-26].

From the speech processing literature, it is evident that psychological and physiological datasets are being solved using multiple machine learning methods. This raises the question of how we could know whether our data are truly ordinal? In other words, which learning approach would be most suitable for a given dataset? Considering its success in many affective computing fields [4, 5, 27], deep learning is a good fit to exploit the nature of ground truth. In this paper, we investigate the problem of choosing the learning approach by proposing a novel way to combine multiple loss functions, evaluated on both synthetic data and six real affective computing datasets.

3. Proposed Framework

The proposed framework (Fig. 1) for assessing the ordinality of data is a neural network system with a special hyper-parameter γ. The parameter γ provides information in an empirical setting on the most suitable analysis method for a given dataset, namely whether a given problem should be treated as classification, ordinal regression, regression or somewhere in between. This is because the learning process in the backend system is expected to be optimized when the loss function (controlled by γ) best fits the data. γ is confined to $\gamma \in [-1, +1]$, progressively changing from *classification* $(\gamma = -1)$ to *ordinal regression* $(\gamma = 0)$ and then to *regression* $(\gamma = +1)$. Intermediate γ values represent the combined states between classification-ordinal regression and ordinal regression-regression.

The proposed loss function (1) is thus a linear combination of three loss functions that is apt for three learning methods $(L_{cls}, L_{regr}, L_{ord})$. The contribution of each loss term is adjusted by three coefficients α, β, $(1-\alpha)(1-\beta)$ such that a maximum of two loss terms are considered in the total loss (L_t) calculation for any given γ value (Fig. 2).

$$L_t = \alpha L_{cls} + (1-\alpha)(1-\beta)L_{ord} + \beta L_{regr}$$
$$\text{where } \alpha = \begin{cases} |\gamma|, & \gamma < 0 \\ 0, & \gamma \geq 0 \end{cases} \text{ and } \beta = \begin{cases} \gamma, & \gamma > 0 \\ 0, & \gamma \leq 0 \end{cases} \quad (1)$$

The classification (L_{cls}) and regression (L_{regr}) loss terms here are selected as binary cross entropy loss (2) and Mean Squared Error (MSE) loss (4) which are well-established in the machine learning literature.

$$L_{cls} = -\sum_{i=1}^{N}\sum_{k=1}^{K} p_{ik} log(\bar{p}_{ik}) + (1-p_{ik}) log(1-\bar{p}_{ik}) \quad (2)$$

$$\text{where } \bar{p}_{ik} = \frac{(K-1)! {p_x}^{k-1} (1-p_x)^{K-k}}{(k-1)!(K-k)!} \quad (3)$$

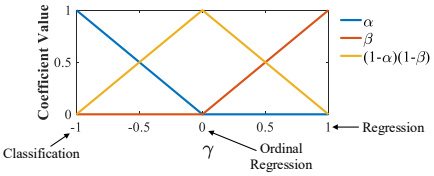

Figure 2: *Variation of the three coefficients of the three loss terms with respect to γ. α and β represent the coefficients for classification and regression loss terms respectively.*

where N is the number of data samples and K is the maximum value in the ordinal scale. p_{ik} is the target probability for k^{th} class and $\bar{p}_{ik}$ is the output probability. In order to calculate output probabilities from a single-dimensional output p_x, it is assumed that the output probabilities follow a binomial distribution (3), as described in [28]. Prior to calculation of $\bar{p}_{ik}$, the *Sigmoid* activation function is applied to the neural network output, s.t. $sigmoid\big(h(x)\big) = p_x \in [0,1]$ (Fig. 1). $h: x \rightarrow y$ is the learning function of the neural network. *ReLU* is applied prior to calculation of L_{regr} calculation.

$$L_{regr} = \sum_{i=1}^{N} \big(y_i - h(x_i)\big)^2 \quad (4)$$

The loss function in PrefNet, which is referred to as pairwise preference loss, is proposed as the ordinal loss term (L_{ord}), and its expression is given by:

$$L_{ord} = \sum_{(a,p)} l_{ap}$$
$$l_{ap} = max\{0, \theta + h(x_a) - h(x_p)\} \quad (5)$$
$$\forall (a,p)\ x_a, x_p \in X \times X \text{ s.t. } y_p > y_a$$

θ is the margin and was set to 0.2 in the following experiments. Among the few differentiable ordinal loss functions, pairwise preference loss is apposite for the proposed system architecture and been previously shown to give good performance [15].

Combined loss functions are not uncommon in the neural network literature. For example, the neural network system presented in [29] used a combination of MSE loss and a ranking loss function based on cross-entropy in order to incorporate the notion of ranking into standard regression network. Similarly, logistic regression loss function along with pairwise hinge loss was applied to ordinal classification in [16]. Therefore, a linear combination of multiple loss functions could introduce different properties to the neural network system. However, to our knowledge, using a parameter to trade off individual loss functions has not been used as an indicator of dataset properties.

4. An Investigation of Ordinal Data Structures using Synthetic Data

To understand the behaviour of the proposed framework, we conducted experiments on synthetic data. Synthetic data permits modification of the structure of the data with known ground truth, and hence can provide a clear expected value for γ. In the machine learning literature, ordinal synthetic datasets have been generated broadly in two ways: (i) as ordered Gaussian clusters [30, 31] (ii) using linear or non-linear equations followed by thresholding to derive ordinal categories [32]. In this work, synthetic datasets were generated according to the former approach for visualisation purposes and because modifying it to generate classification-like and regression-like datasets is also much easier. More specifically, three aspects will be investigated: number of classes (Section 4.2), non-linear structure (Section 4.3), and intra-class spacing (Section 4.4).

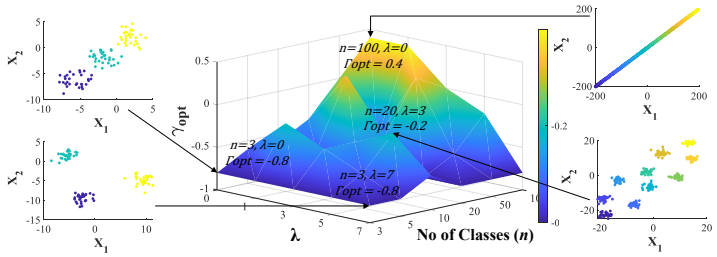

Figure 3: γ_{opt} *as a function of n (number of classes) and λ (mean-shifted magnitude). The four plots of example synthetic datasets correspond to four different coordinates in the x-y plane. When n decreases and λ increases,* γ_{opt} *tends towards* -1*, implying that classification properties become prominent. For large n and small λ,* γ_{opt} *tends towards* 1*, implying regression.*

4.1. Experimental Settings

For experiments on synthetic data we used a neural network consisting of two fully connected layers with layer sizes of 16, 6 and a 1-dimensional output layer. The first two layers adopted the *Leaky-ReLU* activation function. The network was trained using backpropagation with the *Adam* optimiser. The optimum value of γ was selected empirically from among $\gamma \in \{0.2i \mid i \in \{-5, \cdots, 5\}\}$ using Kendall's rank correlation coefficient (Kendall's τ) [33]. Rank correlation measures are preferred for ordinal ground truth [34] as they focus on trend between ground truth and predictions as opposed to metric distance (e.g. RMSE) or ignoring ordering (e.g. classification accuracy). In each experiment, optimum γ, i.e. γ_{opt}, was selected based on the highest average of 5-fold cross validation Kendall's τ. Dataset generation is described for each experiment in turn, since the data generation method varies between experiments.

4.2. Ordinality Variation with Number of Classes

To begin, we investigated the optimal γ as a function of the number of classes (i.e. the size of the ground truth scale). We created a set of synthetic datasets with different number of classes, $n \in \{3, 5, 10, 20, 50, 100\}$. Usually, ordinal datasets consist of 5-10 classes. However, depression scales are much larger in size: e.g. the PHQ-8 [7] scale has 25 levels, and the BDI-II [8] scale has 64. The synthetic dataset was generated as bivariate Gaussian clusters with equally spaced means and 0.5 standard deviation, forming a straight line (Fig. 4). Each class comprised 30 data points. According to Fig. 4, when the scale size is between 5-50, γ_{opt} is closer to 0, suggesting an ordinal structure to the data. On the contrary, when the number of classes is small $(n = 3)$ γ_{opt} is -0.8, implying that it is more of a classification task. Similarly, when the number of classes is large $(n = 100)$, the problem becomes more of a regression problem, i.e. γ_{opt} is 0.6. Ordinal regression models are designed based on the assumption that the ordinal categories are envisaged as contiguous intervals on a latent continuous variable (e.g. POM [21]). We speculate that when the scale size is considerably large, contiguous regions in this continuous scale approximate a line, allowing regression to perform better.

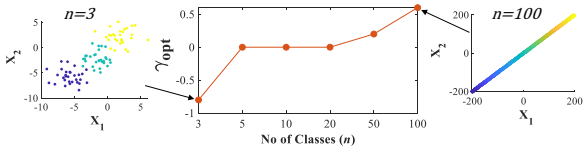

Figure 4: γ_{opt} *as a function of n (number of classes). The leftmost and rightmost figures are the representation of synthetic datasets used in two extreme cases. It is evident that* γ_{opt} *has a positive correlation with n.* $\gamma_{opt} \sim -1$ *is associated with a classification loss function, while* $\gamma_{opt} \sim +1$ *is associated with a regression loss function.*

4.3. Ordinality under Non-Linear Structures

This experiment was designed to differentiate between classification and ordinal regression: i.e. observing the behaviour of γ under a non-linear setting in which the ordinality is distorted. Bivariate Gaussians were generated on a straight line and then shifted in a perpendicular direction. Fig. 3 shows γ_{opt} as a function of $\lambda \in \{0,1,3,5,7\}$, the amount of shifting, repeated for different scale sizes. As would be expected, when the ordering between classes is distorted, γ_{opt} goes towards -1, indicating that nominality has become prominent. The consistent behaviour of γ_{opt} across different scale sizes (Fig. 3) is further encouragement that the expected dominance of classification loss holds when ordering is not preserved.

4.4. Ordinality vs Intra-Class Spacing

Synthetic datasets used in the previous experiments have had uniformly separated Gaussian clusters, however features that fall into uniformly-spaced classes are seldom encountered in real applications. In this section, the behaviour of γ_{opt} is studied under varying spacing between Gaussian clusters. In this experiment, five different scenarios from *equally* spaced to *unequally* spaced clusters were generated, denoted as $I_1, \cdots, I_5$

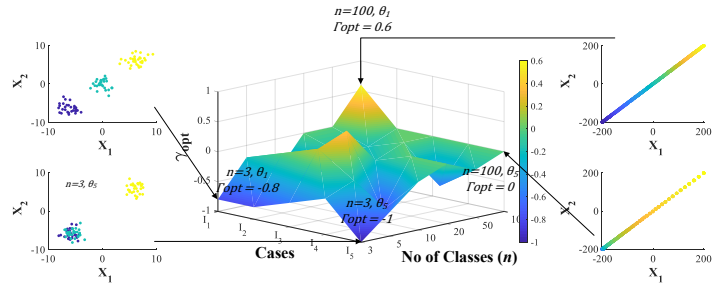

Figure 5: γ_{opt} *as a function of n (number of classes) and cases I representing various non-uniform spacings. The four plots of synthetic data correspond to four extreme examples. When spacing becomes more non-uniform,* γ_{opt} *converges to 0 implying ordinal properties become prominent.*

(Fig. 5). Any of the I configurations (except I_1) represents a slightly distorted class uniformity, and hence γ_{opt} closer to 0 is observed across different I configurations. When $n = 100$ and the Gaussians are uniformly distributed (I_1), the problem is approximately a regression problem, so a higher γ_{opt} is observed (section 4.2). When spacing becomes more and more unbalanced $(I_2 \rightarrow I_5)$, the γ_{opt} converges to 0 implying that an ordinal loss is better suited when the class ordering is highly non-uniform. Throughout the different cases in Fig. 5, γ_{opt} behaves reasonably consistently with expectation.

5. Ordinality of Real Speech Data

While the simulated experiments using the proposed framework offer some interesting insights into choosing a learning approach, it is important to study the γ_{opt} of well-known datasets used in affective computing and related fields.

5.1. Experimental Settings

The AViD [35] corpus and DAIC-WOZ [36] corpora are depression datasets labelled using self-assessment tools: AViD using BDI-II, DAIC-WOZ using PHQ-8. Stroop is a cognitive load corpus in which speakers perform three different reading tasks designed to experience different cognitive loads: *low*, *medium*, *high* [1]. VAM [9], IEMOCAP [10] and RECOLA [11] are emotional datasets, all of which are labelled in arousal and valence dimensions. VAM and IEMOCAP were labelled from 1 – 5 by multiple annotators. VAM provides the average of all annotations as the final rating, and herein the scale was split into five equal sized classes. For IEMOCAP, the median of all annotations was taken as the ground truth. The RECOLA corpus was annotated on a continuous scale between $[-1, +1]$ and herein discretised evenly into two scales: 5 and 100 classes.

The back-end model was as described in Section 3.1 except the DNN was replaced with a LSTM (herein referred as γLSTM), incorporating dynamic temporal information using LSTM is well-established in speech processing [4, 37]. The backend network consists of one LSTM layer followed by two fully connected layers. Model parameters and hyper-parameters were chosen empirically. Front-end feature sets were chosen according to typical baseline features for the dataset.

5.2. Results and Discussion

Table 1 provides γ_{opt} of six real databases, derived based on average Kendall's τ of 5-fold cross validation (column 4). The two depression datasets (AViD and DAIC-WOZ), which have 25-classes and 64-classes, were tested on their originally annotated scale as well as on a 5-class scale regrouped based on the definition of the respective depression scales. It was somewhat unanticipated to observe γ_{opt} closer to 0 even for a larger number of classes (section 4.2). However, it could be that all the classes are off-line (Fig. 3, $n = 100$, $\lambda = 3$) but regrouping into 5-classes may have brought them back to a line. Another possible explanation could be found from Fig. 5: spacing between classes may be non-uniform ($n = 100, I_5$). Nevertheless, four γ_{opt} values in Table 1 suggest applying ordinal regression for depression prediction.

The Stroop (cognitive load) dataset, which is annotated with a 3-class ordinal scale, was tested with two feature sets. We believe the structure of the two feature spaces are different and hence a different γ_{opt} was observed. Therefore, not only the nature of ground truth, even the structure of data may matter in deciding the modelling method. Both γ_{opt} being closer to 0 implies both feature sets are ordinal and hence ordinal modelling algorithms should be applied. On the contrary, three emotional datasets exhibit contrasting behaviours. IEMOCAP, which is usually treated as a classification problem, also reports -1 for γ_{opt}. Even though it is annotated on an ordinal scale, the underlying data structure may not be ordered on a manifold, like $n = 5, \lambda = 7$ in Fig. 3. VAM favours ordinal regression, in keeping with the ordinal self-assessment mannikin annotation method. RECOLA, which has continuous annotation, suggests regression for 100 classes, agreeing with the observations in section 4.2. However, regrouping into five classes apparently transforms the regression-like problem into an ordinal-like problem, suggesting that a change in labelling may have important implications for the choice of learning approach (but not always – see AViD result). The last column in Table 1 presents γ_{opt} calculated using CCC, which is a common evaluation metric in emotion recognition. In most cases there is agreement between the two γ_{opt} values, suggesting a robustness to the evaluation metric.

Table 1: *Optimum γ values of different datasets. $\gamma_{opt} \sim -1$ is associated with a classification loss function, whilst $\gamma_{opt} \sim +1$ is associated with a regression loss function.*

Database	System Description	Scale Size	γ_{opt} (τ)	γ_{opt} (CCC)
DAIC-WOZ	MFCC+ γLSTM	5	0	0
	MFCC+γLSTM	25	-0.4	0.8
AViD	MFCC+γLSTM	5	-0.2	-0.2
	MFCC+γLSTM	64	0.0	-0.2
Stroop	MFCC_SDC + γLSTM	3	0.2	-0.2
	eGeMAPS(lld)+γLSTM	3	-0.4	-0.4
VAM	IS09+γLSTM	5	0.4	0.4
IEMOCAP	IS09+γLSTM	5	-1	-1
RECOLA	MFCC+γLSTM	5	-0.2	-0.4
	MFCC+γLSTM	100	0.8	1

∗ *MFCC: Mel Frequency Cepstral Coefficients, SDC: Shifted Delta Cepstra, IS09: Interspeech 2009 feature set, γLSTM: proposed backend with a LSTM*

6. Conclusion

This paper has presented a novel framework built on neural networks in which a single hyper-parameter γ trades off between classification, regression, and ordinal regression loss functions. Hence, the optimum γ (γ_{opt}) for a particular dataset indicates the suitability of the three loss functions, hinting at the structure of the data and suggesting how we should formulate and evaluate problems on it. The principles of the proposed framework were explored using synthetic datasets: when increasing or decreasing the number of classes on an ordinal scale, the behaviour of γ_{opt} suggests to be treated as a regression or a classification problem respectively. For real ordinal datasets, there may be extreme data structures for which classification or regression is better suited, and the proposed framework could assist with choosing the appropriate modelling approach. Application of the proposed framework to six affective computing speech datasets suggests that more focus should be placed on ordinal regression instead of the conventional approaches of classification and regression.

It is acknowledged that one limitation of this work is that simulated experiments presented in Section 4 do not provide full coverage of all possible conditions, which will be left for future work. Future work also includes introducing a learnable γ to the proposed framework to avoid exhaustive evaluation of different γ values.

7. References

[1] B. Yin, F. Chen, N. Ruiz, and E. Ambikairajah, "Speech-based cognitive load monitoring system," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008: IEEE, pp. 2041-2044.

[2] A. C. Trevino, T. F. Quatieri, and N. Malyska, "Phonologically-based biomarkers for major depressive disorder," *EURASIP Journal on Advances in Signal Processing,* vol. 2011, no. 1, p. 42, 2011.

[3] Y. Kim and E. M. Provost, "Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013: IEEE, pp. 3677-3681.

[4] J. Zhao, R. Li, S. Chen, and Q. Jin, "Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, 2018, pp. 65-72.

[5] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, and R. Garg, "Multi-level Attention network using text, audio and video for Depression Prediction," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 81-88.

[6] L. Zhang, J. Driscol, X. Chen, and R. Hosseini Ghomi, "Evaluating Acoustic and Linguistic Features of Detecting Depression Sub-Challenge Dataset," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 47-53.

[7] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *Journal of Affective Disorders,* vol. 114, no. 1, pp. 163-173, 2009.

[8] A. T. Beck, R. A. Steer, and G. K. Brown, "Beck depression inventory-II," *San Antonio,* vol. 78, no. 2, pp. 490-498, 1996.

[9] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *2008 IEEE international conference on multimedia and expo*, 2008: IEEE, pp. 865-868.

[10] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation,* vol. 42, no. 4, p. 335, 2008.

[11] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, 2013: IEEE, pp. 1-8.

[12] G. N. Yannakakis, R. Cowie, and C. Busso, "The Ordinal Nature of Emotions: An Emerging Approach," *IEEE Trans. on Affective Computing,* 2018.

[13] S. Parthasarathy, R. Cowie, and C. Busso, "Using agreement on direction of change to build rank-based emotion classifiers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 24, no. 11, pp. 2108-2121, 2016.

[14] T. F. Yap, J. Epps, E. Ambikairajah, and E. H. Choi, "Voice source under cognitive load: Effects and classification," *Speech Communication,* vol. 72, pp. 74-95, 2015.

[15] H. P. Martinez, G. N. Yannakakis, and J. Hallam, "Don't classify ratings of affect; rank them!," *IEEE Trans. on Affective Computing,* vol. 5, no. 3, pp. 314-326, 2014.

[16] Y. Liu, A. Wai Kin Kong, and C. Keong Goh, "A constrained deep neural network for ordinal regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 831-839.

[17] F. Fernández-Navarro, A. Riccardi, and S. Carloni, "Ordinal neural networks without iterative tuning," *IEEE transactions on neural networks and learning systems,* vol. 25, no. 11, pp. 2075-2085, 2014.

[18] N. Cummins, V. Sethu, J. Epps, J. R. Williamson, T. F. Quatieri, and J. Krajewski, "Generalized Two-Stage Rank Regression Framework for Depression Score Prediction from Speech," *IEEE Trans. on Affective Computing,* 2017.

[19] R. Lotfian and C. Busso, "Practical considerations on the use of preference learning for ranking emotional speech," in *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2016: IEEE, pp. 5205-5209.

[20] S. Jayawardena, J. Epps, and E. Ambikairajah, "Support Vector Ordinal Regression for Depression Severity Prediction," in *International Conference on Information and Automation for Sustainability (ICIAFS)*, Sri Lanka, 2018: IEEE.

[21] P. McCullagh, "Regression models for ordinal data," *Journal of the Royal Statistical Society. Series B (Methodological),* pp. 109-142, 1980.

[22] V. E. Farrugia, H. P. Martínez, and G. N. Yannakakis, "The preference learning toolbox," *arXiv preprint arXiv:1506.01709,* 2015.

[23] C. Burges *et al.*, "Learning to rank using gradient descent," in *Proc. of the Intl. Conf. on Machine Learning*, 2005: ACM, pp. 89-96.

[24] J. Cheng, Z. Wang, and G. Pollastri, "A neural network approach to ordinal regression," in *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, 2008: IEEE, pp. 1279-1284.

[25] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output cnn for age estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4920-4928.

[26] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2002-2011.

[27] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli, "Multimodal Measurement of Depression Using Deep Learning Models," in *Proc. of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017: ACM, pp. 53-59.

[28] J. P. Da Costa and J. S. Cardoso, "Classification of ordinal data using neural networks," in *European Conference on Machine Learning*, 2005: Springer, pp. 690-697.

[29] W. Han, H. Li, F. Eyben, L. Ma, J. Sun, and B. Schuller, "Preserving actual dynamic trend of emotion in dimensional speech emotion recognition," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, 2012: ACM, pp. 523-528.

[30] B.-Y. Sun, J. Li, D. D. Wu, X.-M. Zhang, and W.-B. Li, "Kernel discriminant learning for ordinal regression," *IEEE Trans. on Knowledge and Data Engineering,* vol. 22, no. 6, pp. 906-910, 2010.

[31] W. Waegeman, B. De Baets, and L. Boullart, "ROC analysis in ordinal regression learning," *Pattern Recognition Letters,* vol. 29, no. 1, pp. 1-9, 2008.

[32] R. Herbrich, T. Graepel, P. Bollmann-Sdorra, and K. Obermayer, "Learning preference relations for information retrieval," in *ICML-98 Workshop: text categorization and machine learning*, 1998, pp. 80-84.

[33] M. G. Kendall, "A new measure of rank correlation," *Biometrika,* vol. 30, no. 1/2, pp. 81-93, 1938.

[34] S. Jayawardena, J. Epps, and E. Ambikairajah, "Evaluation Measures for Depression Prediction and Affective Computing," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019: IEEE, pp. 6610-6614.

[35] M. Valstar *et al.*, "AVEC 2013: the continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 2013: ACM, pp. 3-10.

[36] J. Gratch *et al.*, "The Distress Analysis Interview Corpus of human and computer interviews," in *LREC*, 2014: Citeseer, pp. 3123-3128.

[37] L. Chao, J. Tao, M. Yang, and Y. Li, "Multi task sequence learning for depression scale prediction from video," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, 2015: IEEE, pp. 526-531.