A Study on Affect Model Validity: Nominal vs Ordinal Labels

David Melhart DAVID.MELHART@UM.EDU.MT
Konstantinos Sfikas KON.SFIK@GMAIL.COM
Institute of Digital Games - University of Malta, Malta

Giorgos Giannakakis GGIAN@ICS.FORTH.GR
Computational BioMedicine Laboratory - Foundation for Research and Technology Hellas, Greece

Georgios N. Yannakakis GEORGIOS.YANNAKAKIS@UM.EDU.MT
Antonios Liapis ANTONIOS.LIAPIS@UM.EDU.MT
Institute of Digital Games - University of Malta, Malta

Abstract

The question of representing emotion computationally remains largely unanswered: popular approaches require annotators to assign a magnitude (or a class) of some emotional dimension, while an alternative is to focus on the relationship between two or more options. Recent evidence in affective computing suggests that following a methodology of ordinal annotations and processing leads to better reliability and validity of the model. This paper compares the generality of classification methods versus preference learning methods in predicting the levels of arousal in two widely used affective datasets. Findings of this initial study further validate the hypothesis that approaching affect labels as ordinal data and building models via preference learning yields models of better validity.

Keywords: preference learning, classification, support vector machines, model evaluation, affect modelling

1. Introduction

Capturing and reliably predicting the nuances of changing emotional states is a central problem of affective computing. The issue is complex, encompassing not only the experimental protocol for data collection but also labelling to processing emotional data. Despite readily available tools such as Likert scales (Likert, 1932) or the Self-Assessment Manikin (Morris, 1995), labelling emotions remains a challenge. Standard methods of collecting and processing annotations often rely on *absolute* ratings which are processed as scalar or converted to nominal values (Allen and Seaman, 2007; Yannakakis and Martínez, 2015). However, processing ratings as scalar or nominal values leads to a range of issues in representing, measuring, understanding, and modelling affect (Yannakakis et al., 2017). There is ample evidence to suggest that decision making and emotional processing rely on *anchoring-biases* (Damasio, 1994; Seymour and McClure, 2008) and are subject to *adaptation effects* (Helson, 1964). This ever-changing baseline, to which we compare new experiences, means we treat information as ordinal and our evaluation is subjective. Doing so has two main effects: first, it is easier for us to pick the better option of two or more outcomes than to assign an absolute value to an emotion (Yannakakis et al., 2017). Second, asking people to rate or classify their experiences may introduce biases as annotators have to interpret the provided scales and rating systems, which lead to inconsistencies and unreliable reporting (Yannakakis and Martínez, 2015).

Motivated by the developing evidence for the advantages of the ordinal labelling approach (Yannakakis et al., 2017), this paper compares preference learning to classification across two popular datasets within affective computing: the DEAP and the AMIGOS datasets. To compare two types of algorithms on these datasets, we first convert affect ratings to classes and preferences and then test the models' accuracy in predicting a unseen (validation) set of ratings. For fairness in comparisons, both approaches use support vector machines (SVMs) as the underlying methodology. Our key results reveal that preference learning is a preferred method for constructing models of arousal in the examined datasets, as it yields more general models compared to the ones built via classification.

This paper contributes to the evidence that better models of affect could be built from existing datasets in affective computing if affect labels are treated as ordinal data and models are built via preference learning. We revisit the line of work behind preference annotations (Martinez et al., 2014; Yannakakis and Martínez, 2015; Yannakakis et al., 2017) and adapt these findings to robust machine learning techniques (SVMs) and state-of-art datasets (Koelstra et al., 2012; Miranda-Correa et al., 2017) with more granular affect ratings than traditional Likert-like scales. Our key hypothesis is that ranking approaches yield higher cross-validation accuracies than corresponding classification algorithms. We also argue for the use of preference learning on datasets built for classification with granular ratings, as it affords more precision in defining category boundaries. It should be noted that we test this hypothesis in real-world datasets, unlike work by Martinez et al. (2014) on synthetic data.

2. Background

In this section we outline the Classification and Preference Learning paradigms that are compared for their predictive capacity in the selected datasets.

2.1. Classification for Affect Modelling

Classification (CL) is the supervised machine learning technique where a predictive model classifies the provided data points into discrete categories. More formally, every instance $X = [x_i | i = 0 \dots n]$ is assigned a discrete label from a set of predefined set $L = [\lambda_j | j = 0 \dots k]$. The algorithm learns a model which predicts the label $\lambda_j \in L$ for each new datapoint provided. Our experiments use a binary classifier. CL is a reliable and widely used method in affective computing (Kapoor, 2015) and performs very well in user-dependent affect modelling (Al Zoubi et al., 2012). However, in user-independent affect detection CL is challenged as it reduces a model's possible output into a set of finite and discrete states; this introduces a bias when determining the split criterion between categories (Martinez et al., 2014). Moreover, the stark separation between classes hides the granularity and ordinal representation of emotions (Yannakakis et al., 2017). This paper uses Support Vector Classifiers (SVC) as our CL baseline for comparisons with rank-based SVMs.

2.2. Preference Learning for Affect Modelling

Preference learning (PL) is a supervised learning technique where a learned model predicts the preference order either explicitly between two datapoints at a time, or implicitly by applying multi-label classification (Fürnkranz and Hüllermeier, 2003). Pairwise PL exploits one-to-one classification (Fürnkranz and Hüllermeier, 2003) to create an explicit ranked preference between two datapoints. While CL treats the output as nominal labels, in PL the output is a rank order. More formally, we assign a label $L = [\lambda_i | i = 0 \ldots n]$ for every instance $X = [x_i | i = 0 \ldots n]$ and provide the learning algorithm with the preferred order for each pair of $(x_i, x_j) \in X$ in the form of their corresponding labels $\lambda_i \succ \lambda_j$, where λ_i is preferred over λ_j. This list of preferences P is a subset of all possible rankings ($P \subseteq L \times L$) as in some cases no clear order can be inferred. The PL algorithm aims to learn a predictive preference model between any two instances of the provided dataset.

Unlike classification, learning affect preferences retains information on their underlying order, revealing global and local preference relations. Although nominal values cannot always be processed via PL (if there is no inherent order), ratings can easily be converted into ranks for PL purposes (Yannakakis et al., 2017). Previous studies on PL methods in affective computing focused on labelling data (Yannakakis and Martínez, 2015) and processing continuous annotations into ordinal data for general affect modelling (Camilleri et al., 2017), Instead, this paper extends the work of Martinez et al. (2014) on comparing PL to CL methods, focusing on SVMs testing the performance of PL and CL on popular affective datasets. Our aim is to generalise the findings of an ordinal labelling and processing approach (Yannakakis et al., 2017) to new machine learning methods and datasets, and lay down the the groundwork for future studies focusing on general models of affect.

3. Experiment

Our experiment compares models trained via classification and preference learning (see Section 3.2) on two popular datasets (see Section 3.1). Our machine learning algorithms are built on support vector machines that use a radial basis function kernel (RBF): we use support vector classifiers (SVC) for classification, and ranking support vector machine (rankSVM) for preference learning. The paper uses the implementation of rankSVM available in the Preference Learning Toolbox (Farrugia et al., 2015) which is based on the original rankSVM algorithm (Joachims, 2002). This ranking method is based on a pairwise approach: the algorithm approximates a binary classifier not for the whole dataset but for each provided pairwise comparison and learns to pick the preferred instance.

3.1. Datasets

In this paper we use two publicly available datasets, the *Database for Emotion Analysis; using Physiological signals* (DEAP) (Koelstra et al., 2012) and *A dataset for Mood, personality and affect research on Individuals and GrOupS* (AMIGOS) short video dataset (Miranda-Correa et al., 2017). Both sets use short videos as elicitors of emotion and track a wide array of physiological signals and have similar annotation techniques. Motivated by their demonstrated links to arousal, this paper uses heart rate variability (HR) and skin conductance (SC) signals of those datasets as input for our models: (1) average HR, (2) standard deviation of the HR normal-to-normal (NN) interval, (3) root mean square of successive differences in HR, (4) HR NN intervals that differ more than 50ms, (5) average SC level, (6) number of significant SC responses, (7) sum of the amplitudes of significant SC responses. The output of our models are the annotated arousal ratings, which were provided through a Self-Assessment Manikin (SAM) in both datasets, as a floating-point scale within $[1, 9]$. The DEAP dataset provides observations on 23 annotators, each with 40 feature sets and corresponding ratings, totalling 920 data points. The AMIGOS dataset has data from 40 annotators, each with 16 feature sets and ratings, totalling 640 data points.

3.2. Experimental Protocol

Experiments in this paper revolve around leave-one-annotator out cross-validation. As a preliminary step, we remove one participant (randomly chosen) from each dataset, without applying any transformation to their data: this participant's raw ratings later to validate and compare the performance of the SVCs and rankSVMs. We pre-process the data of the remaining 22 and 39 participants (for DEAP and AMIGOS respectively), first by converting their annotated ratings into binary classes (for CL) and pairwise comparisons (for PL).

For classification we use two threshold options to create binary classes. In the first case, classes are split between low arousal for SAM scores of $[1, 5)$ and high arousal of $(5, 9]$. We identify this setup as SVC with a class boundary of 5, i.e. SVC(5). SVC(5) gives us 871 classified instances out of the 920 in the DEAP dataset and 550 out of 640 in the AMIGOS dataset. The second case identified as SVC(4.5-5.5) uses a broader boundary, splitting data between low arousal for SAM scores of $[1, 4.5)$ and high arousal of $(5.5, 9]$. SVC(4.5-5.5) has 749 classified items in the DEAP and 505 in the AMIGOS dataset. We use these class thresholds because any other splitting criterion of high versus low arousal categories results in substantially smaller datasets that are not directly comparable to the dataset produced by pairwise preferences.

For preference learning, we create a list of the pairwise comparisons of temporally adjacent instances in the form of $\lambda_i \succ \lambda_j$ (λ_i is preferred over λ_j). Similarly to SVCs, we use two different configurations to construct this list. The first one considers any difference between two ratings as a clear ranking and is identified as rankSVM(0); the second treats trivial differences (absolute difference below 0.1) as equal rank and is identified as rankSVM(0.1). From the possible 919 comparisons for DEAP, rankSVM(0) yields 842 pairs and rankSVM(0.1) yields 760 pairs; for the 639 possible comparisons for AMIGOS, rankSVM(0) yields 509 pairs and rankSVM(0.1) yields 475 pairs.

In the next step, we divide these participants into cross-validation folds, one per participant (save for the one participant we reserved in the preliminary step). Within these folds, we normalise the data and balance them across participants with oversampling. After oversampling, we train each fold on 840 class instances (in case of SVCs) and 819 pairs of pairwise preferences (in case of rankSVMs) on the DEAP dataset and 608 class instances and 532 pairwise preferences on the AMIGIOS dataset. To find the best parameter for the $RBF\gamma$, we use sensitivity analysis by tuning the parameter between 10 and 100 in increments of 10 (see Figure 1). We tune the parameters for each algorithm independently and choose the parameters that result in the best average prediction accuracy on the test set (i.e. one participant, cross-validated across all folds).

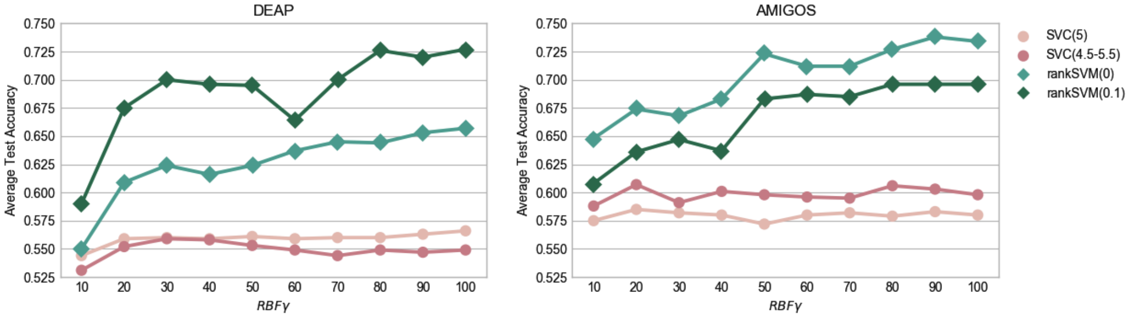

Figure 1: Sensitivity analysis of the $RBF\gamma$ parameter on SVCs and rankSVMs for DEAP (left) and AMIGOS (right) datasets. The plot shows the average accuracy on the test sets.

It should be noted that the test accuracy is not directly comparable between algorithms due to differences at both the algorithmic and the dataset level (e.g. the number of data points are close but not identical among the four setups of Fig. 1). For a fair comparison, SVCs and rankSVMs are compared based on the trained models' predictions of the one participant's data which we reserved in the preliminary step. Inspired by comparisons performed by Lotfian and Busso (2016) and Martinez et al. (2014), we evaluate SVCs (for CL) and rankSVMs (for PL) by building a global order based on the predicted variables' relative distance to the decision boundary and comparing it to the global order of the participant's raw ratings with Kendall's τ.

As the RBF kernel maps our input space to a feature space with infinite dimensions, we cannot calculate the absolute distance from the boundary. However, we are able to use the decision function as a relative approximation. For this purpose, we use Equation 1, where x_i and y_i are the input and output of the test set, x' is the input to be predicted, α_i is a coefficient which separates support vectors from the rest of the datapoints, b is the bias of the model, and K is the RBF kernel (Equation 2).

$$f(x') = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x') + b \tag{1}$$

$$K(x, x') = e^{-\gamma \|x - x'\|^2} \tag{2}$$

As both SVC and rankSVM use the same decision function to evaluate this distance, this metric provides us with a common ground across different SVM implementations. In both cases a higher distance to the boundary equates to a higher confidence in the classification or ranking of the instance. Sorting based on this distance, a high Kendall's τ with the participant's global order of annotated preferences (i.e. the most arousing video clip ranked first, based on the participant's SAM annotation) means that a model closely matches the participant's affect preferences.

4. Results

In this section we discuss the key results of performance comparisons between each SVM type across the two datasets examined. We report each algorithms' performance per dataset for the best $RBF\gamma$ value found via the sensitivity analysis shown in Fig. 1.

Table 1: Performance of SVCs and RankSVMs on the DEAP and AMIGOS datasets. Results show the cross-validation accuracies (along with the chance-based baseline) and the Kendall's τ calculated as per Section 3.2.

	DEAP				AMIGOS			
		Accuracy		Kendall's τ		Accuracy		Kendall's τ
Algorithm	$RBF\gamma$	Base	Avg (Max)	Avg (Max)	$RBF\gamma$	Base	Avg (Max)	Avg (Max)
SVC(5)	100	0.56	0.57 (0.83)	0.14 (0.22)	30	0.58	0.58 (1.00)	0.15 (0.28)
SVC(4.5-5.5)	40	0.55	0.56 (0.88)	0.15 (0.24)	20	0.60	0.61 (1.00)	0.13 (0.23)
rankSVM(0)	100	0.51	0.66 (0.77)	0.17 (0.26)	90	0.53	0.74 (0.93)	0.10 (0.37)
rankSVM(0.1)	100	0.50	0.73 (0.85)	0.19 (0.26)	90	0.53	0.70 (0.93)	0.13 (0.40)

4.1. DEAP

During training, rankSVMs perform significantly, exceeding their chance-level baseline accuracies, while SVCs can barely improve on their respective baseline. The cross-validation results in terms of accuracy are shown in Table 1, along with the respective baseline accuracy (i.e. always guessing the most common rank or label). While accuracies across different machine learning tasks and dataset splits are not directly comparable, it is evident that the average accuracy improves in a much more pronounced manner over its baseline value for rankSVMs. For the purposes of comparison, we use the Kendall's τ between the scoring of each prediction from the decision function of the SVM at hand and the raw absolute ratings as they were recorded by the annotators. Based on τ scores of Table 1), both rankSVMs clearly outperform the SVCs at capturing a more general global order and predict an unseen order, especially for rankSVM(0.1).

4.2. AMIGOS

Results obtained from AMIGOS, reported in Table 1, show similar patterns to the DEAP dataset. Although the baselines in this dataset are less balanced, we are accounting for this issue by applying weights to the classes while training the SVCs. Even so, SVCs grossly underperform compared to the rankSVMs: this is especially true for other $RBF\gamma$ values (see Fig. 1) outside the best ones reported in Table 1. It is evident that SVCs barely surpass the baseline on average, while high maximum accuracies may point to over-fitting. Directly comparing the two algorithms through the Kendall's τ validation score shows a similar picture, with rankSVMs outperforming the SVCs. Interestingly, the SVCs perform quite well, on par—and in case of SCV(5)—beyond the ranksSVMs. However, seeing how its test accuracy is only at or around the baseline, this could be attributed to the distribution of the validation set. In contrast, rankSVMs perform remarkably well on the test—which suggests that (unlike SVCs) they do not overfit—and achieve very high maximum values on the validation. Indeed, the best rankSVM(0) is 23% more accurate than its respective baseline, and its maximum Kendall's τ 21% above the best SVC(5).

5. Discussion

This initial study tested the hypothesis that a preference learning method will be able to generalise better than classification when predicting the output of unseen arousal ratings.

Results across two datasets and different splitting criteria reveal that rankSVM performs consistently better on the unseen validation sets than SVC. Based on these observations we can safely validate our hypothesis and provide further evidence for the benefits of representing and processing emotion in an ordinal fashion (Yannakakis et al., 2017).

A major limitation of our study lies in the examined datasets themselves, as each feature set captures the entire session of a short video. In DEAP and AMIGOS the data reflect a discrete judgement and not a continuous change of the participants' evaluation, therefore it is hard to tell if a rating truly reflects a baseline to which participants' measure the next experience. We believe that a more nuanced dataset, where participants' annotation are recorded continuously in time, would improve the performance of preference learning, especially because such data could be processed in a number of different ways (Camilleri et al., 2017). Another limitation is that the study focused on recreating and verifying a basic hypothesis inferred from previous works rather than creating truly general models.

Inspired by Camilleri et al. (2017), future work will investigate the ability of the models to generalize across dataset. As a first step, this involves work with our existing datasets to find the best protocol and metrics for evaluation of the different algorithms. Further, we will explore other user modalities beyond physiology, which may impact the performance of the models. In the future we will be focusing more on datasets with temporally continuous annotation, which provide a more fertile ground for preference learning.

6. Conclusion

This paper documented a comparative study towards obtaining reliable and general affect models. The performance of preference learning (rankSVM) was compared against classification (SVCs) across two popular datasets of affect: DEAP and AMIGOS. Results of this study show the benefits of using preference learning for models of higher validity, and contribute to existing evidence suggesting that ordinal annotation and ordinal processing is a robust way to assess people's emotional and decision making processes, as it better reflects the anchoring mechanisms of our cognition (Yannakakis et al., 2017).

Acknowledgments

This research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 693150.

References

Omar Al Zoubi, Sidney K. D'Mello, and Rafael A. Calvo. Detecting naturalistic expressions of nonbasic affect using physiological signals. *IEEE Transactions on Affective Computing*, 3(3):298–310, 2012.

I. Elaine Allen and Christopher A. Seaman. Likert scales and data analyses. *Quality progress*, 40(7):64, 2007.

Elizabeth Camilleri, Georgios N. Yannakakis, and Antonios Liapis. Towards general models of player affect. In *Intl. Conference on Affective Computing and Intelligent Interaction*, 2017.

Antonio R. Damasio. *Descartes error: Emotion, rationality and the human brain.* Putnam Publishing, 1994.

Vincent E Farrugia, Héctor P. Martínez, and Georgios N Yannakakis. The preference learning toolbox. *arXiv preprint arXiv:1506.01709*, 2015.

Johannes Fürnkranz and Eyke Hüllermeier. Pairwise preference learning and ranking. In *European conference on machine learning*, pages 145–156. Springer, 2003.

Harry Helson. *Adaptation-level theory: an experimental and systematic approach to behavior.* Harper, 1964.

Thorsten Joachims. Optimizing search engines using clickthrough data. In *Intl.Conference on Knowledge Discovery and Data Mining*, pages 133–142. ACM, 2002.

Ashish Kapoor. Machine learning for affective computing: Challenges and opportunities. *The Oxford Handbook of Affective Computing*, page 406, 2015.

Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

Reza Lotfian and Carlos Busso. Practical considerations on the use of preference learning for ranking emotional speech. In *Intl. Conference on Acoustics, Speech and Signal Processing*, 2016.

Hector P. Martinez, Georgios N. Yannakakis, and John Hallam. Dont classify ratings of affect; rank them! *IEEE Transactions on Affective Computing*, 5(3):314–326, 2014.

Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. Amigos: A dataset for mood, personality and affect research on individuals and groups. *arXiv preprint arXiv:1702.02510*, 2017.

Jon D. Morris. Observations: Sam: the self-assessment manikin; an efficient cross-cultural measurement of emotional response. *Journal of advertising research*, 35(6):63–68, 1995.

Ben Seymour and Samuel M. McClure. Anchors, scales and the relative coding of value in the brain. *Current opinion in neurobiology*, 18(2):173–178, 2008.

Georgios N. Yannakakis and Héctor P Martínez. Ratings are overrated! *Frontiers in ICT*, 2:13, 2015.

Georgios N. Yannakakis, Roddy Cowie, and Carlos Busso. The ordinal nature of emotions. In *Intl. Conference on Affective Computing and Intelligent Interaction*, 2017.