ESTIMATION OF ORDINAL APPROACH-AVOIDANCE LABELS IN DYADIC INTERACTIONS: ORDINAL LOGISTIC REGRESSION APPROACH

Viktor Rozgić[1], *Bo Xiao*[1], *Athanasios Katsamanis*[1]
Brian Baucom[2], *Panayiotis G. Georgiou*[1], *Shrikanth Narayanan*[1,2]

[1]Signal Analysis and Interpretation Laboratory (SAIL), Los Angeles, CA, USA
[2]Department of Psychology, University of Southern California, Los Angeles, CA, USA

http://sail.usc.edu[1], baucom@usc.edu[2]

ABSTRACT

Behavioral Signal Processing aims at automating behavioral coding schemes such as those prevalent in psychology and mental health research. This paper describes a method to automatically quantify the approach-and-avoidance (AA) behavior, described by ordinal labels manually assigned by experts using either video-only or video-with-audio. We propose a novel ordinal regression (OR) algorithm and its hidden Markov model (HMM) extension for estimation of AA labels from visual motion capture based and acoustic features. The proposed algorithm transforms the OR to multiple binary classification problems, solves them by independent score-outputting classifiers and fits the cumulative logit logistic regression model with proportional odds (CLLRMP) to vectors of the classifier scores. The time series extension treats labels as states of the HMM with a likelihood function derived from the probabilistic CLLRMP output. We compare performances of the proposed algorithm applying the weighted binary SVMs in the second step (SVM-OLR), its time-series extension (HMM-SVM-OLR) and the baseline multi-class SVM. On the used dyadic interaction dataset the HMM-SVM-OLR achieves the highest estimation accuracies $71.6\ \%$ and $65.7\ \%$ for AA labels assigned respectively using video-only and video-with-audio.

Index Terms— approach-avoidance behavior, dyadic interaction, ordinal regression, cumulative logit model

1. INTRODUCTION

Analysis of behavior patterns in human interpersonal interactions have been in the research focus of psychologists for a long time. For example, coding schemes such as the Couple Interaction Rating System (CIRS) [1] and Rapid Marital Interaction Coding Scheme (RMICS) [2] have been developed for annotation of couples (dyadic) interactions in marital therapy. These and similar schemes define sets of low-level verbal and non-verbal cues of interest (gaze, body orientation, turn taking patterns etc.), rules for deriving intermediate-level behavior codes (e.g. approach-avoidance (AA) codes) by interpretation of low-level signal cues, and additional rules for deriving high-level behavior codes for more complex behaviors (e.g., acceptance, blame levels exhibited by the participants etc.) from low-level cues and intermediate-level behavior codes. The AA behavior code, as an intermediate-level code, is particularly interesting from the engineering perspective. Rules for this code rely strongly on the low-level non-verbal cues, the automatic recognition of which represents a very active research topic [3]. Additionally, the AA behavior is related to a more complex phenomena: emotion and motivation[4] and emotional expressions [5]. Computational modeling of complex human behavior opens up a number challenges, and opportunities, for signal processing. This paper addresses some of them. In our previous work [6], we introduced the multimodal dyadic interaction database and used it for analysis of relations between various non-verbal features and AA labels as defined by psychologists [1]. These labels belong to the ordered set of nine categories, ranging from complete avoidance to complete approach. In this paper we address the estimation of the ordinal AA labels for the same dyadic interactions using the low-level non-verbal signal features. These features represent the basic statistics (mean, minimum, maximum, standard deviation, skewness and kurtosis), of the various video (body orientation, head orientation, hand movement, measure of how opened the postures are) and audio (pitch, energy) based measurements calculated on feature processing window.

In order to address the ordinal nature of the AA labels we propose a universally applicable ordinal regression algorithm that consists of three main steps: (1) we transform the ordinal regression problem to multiple binary classification problems defined by the label ordering; (2) we solve the binary classification problems independently using any classification method that outputs (possibly non-probabilistic) classification score; (3) we fit the cumulative logit logistic regression model with proportional odds (CLLRMP) [7] on vectors obtained by concatenation of scores from the binary classifiers. Since we use continuous features without missing values for estimation of AA categories, we choose to apply weighted binary SVM classifiers with native non-probabilistic scores. Additionally, we propose a simple extension of the proposed algorithm applicable to the time series of ordinal labels. In the extended algorithm, we model the label sequence using a hidden Markov model with a likelihood function based on the probabilistic CLLRMP output.

The two-step ordinal regression algorithm [8] is similar to the method we propose. Methods share the binary classifier reformulation in the first step, in the second step [8] employs probabilistic binary classifiers to directly estimate the cumulative label distribution. However, binary classifiers are trained independently and there is no guarantee that the estimate is monotonically non-decreasing.

We evaluate the proposed estimation methods using leave-part-of-one-session out cross-validation. We present evaluation results for 4 experiments: (1) analysis of dependency between the estimation accuracy and lengths of windows used to calculate the feature statistics; (2) comparison of average estimation accuracies for proposed estimators and the baseline multi-class SVM and analysis of variability in estimation accuracy for different sessions; (3) analysis of confusion matrix differences between estimators; (4) comparison of estimation accuracies for the SVM-OLR and estimator obtained by fitting CLLRMP directly on the original feature vectors.

2. PROPOSED ALGORITHM FOR ESTIMATION OF ORDINAL LABELS

In Section 2.1, we present the transformation of the ordinal regression problem to multiple binary classification problems. In Section 2.2 we present the CLLRMP. We discuss differences when fitting CLLRMP to classifier scores and original feature vectors in Section 2.3. The extension to time series data is presented in Section 2.4.

2.1. Label ordering inspired collection of binary classifiers

Let us introduce the notation used in our paper. We assume that the feature vectors y take values from space $\mathcal{Y}$, and that the ordinal labels o belong to the set O. For simplicity, we denote elements of O as integers $O = \{1, 2, \ldots, K\}$.

We map each ordinal categorical label o to a vector of $K-1$ binary indicators $b(o) = (b_1(o), \ldots, b_{K-1}(o))$ in a way that k^{th} indicator b_k takes value 1 if $o \in \{1, \ldots, k\}$ and value 0 if $o \in \{k+1, \ldots, K\}$. In other words, the described mapping is a redundant, label-ordering inspired, error correcting code.

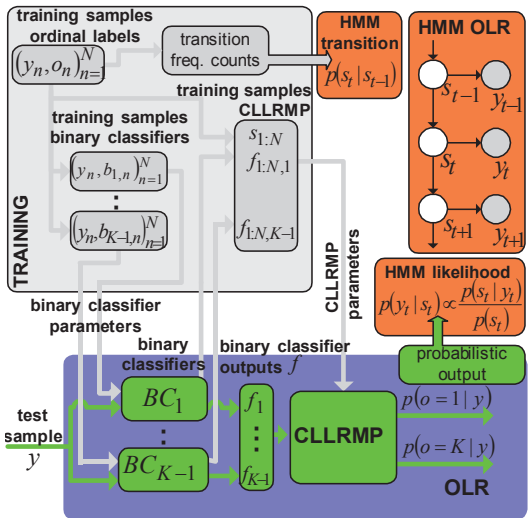

Fig. 1. Proposed method and HMM extension.

We transform the original training dataset $Y = \{(y_n, o_n)\}_{n=1}^N$ to $K-1$ datasets $Y_k = \{(y_n, b_{k,n})\}_{n=1}^N$ $(k = 1, \ldots, K-1)$ with binary labels defined by the described mapping, $b_{k,n} = b_k(o_n)$. On each dataset Y_k we train a binary classifier BC_k.

The collection of trained classifiers $(BC_1, \ldots, BC_{K-1})$ map every feature vector y_n in the training set into the vector of classifier scores $f_n = [f_{1,n}, \ldots, f_{K-1,n}]^T$. Therefore, we say that the collection of trained binary classifiers map Y to $F = \{(f_n, o_n)\}_{n=1}^N$.

2.2. Cumulative logit logistic regression with proportional odds

We fit the CLLRMP [7] to the dataset $F = \{(f_n, o_n)\}_{n=1}^N$ obtained in the previous step. Intuitively, the CLLRMP approximates logits $(logit(x) = log\frac{x}{1-x})$ of the cumulative label distributions by linear functions, with equal slope, of the input vectors (in this case - score vectors f). Formally, the CLLRMP is defined by:

$$\ln \frac{p(o \in \{1, \ldots, k\}|f)}{p(s \in \{k+1, \ldots, K\}|f)} = w_{0,k} + w^T f, \qquad (1)$$

for $k = 1, \ldots, K-1$ where the model parameters $w_0 = [w_{0,1} \ldots w_{0,K-1}]^T$ and $w = [w_1 \ldots w_{K-1}]^T$ are respectively intercept and slope coefficients. The optimal values of the model parameters can be learned from the dataset F in the maximum likelihood sense [9].

An important property of the CLLRMP is that it imposes the stochastic ordering of labels corresponding to different input vectors [7]. This means that it is possible to compare values of the cumulative distribution functions of labels for different score vectors f. This property is summarized by the following equation that follows trivially from Equation 1:

$$p(o = i|f) = p(o = j|f + \frac{w_{0,i} - w_{0,j}}{K-1} w^{-1}), \qquad (2)$$

where w^{-1} is vector of inverse slope coefficients. In the following section we discuss importance of the stochastic ordering in the case when we use weighted binary SVMs in the second step of the proposed estimation method.

2.3. Fitting CLLRMP on classifier score vectors

Let us briefly discuss the implications of the stochastic ordering property imposed by the CLLRMP when it takes the classifier scores as input. The binary classifiers map the original feature vector y to the score vector f whose coordinates have a clear interpretation: if the label of y belongs to $\{k+1, \ldots, K\}$ $(\{1, \ldots, k\})$, then f_k takes low(higher) values. Assuming that classifiers can successfully solve binary classification tasks, there exists a label induced partition of the space of f whose elements are convex sets. Assuming that we "move" f along the line connecting arbitrary $f_1, f_2 \in \mathcal{F}$, it is desirable to have a model such that changes in the cumulative distributions of labels $s(f)$ conditioned on f reflect intersections that f makes with the label induced boundaries. This is exactly the stochastic ordering property. The CLLRMP should fit the classifiers score vectors f better than the original feature vectors y since elements of the label induced partition on $\mathcal{Y}$ are not necessarily convex.

2.4. Hidden Markov model with OLR based likelihood

To address dependencies between consecutive labels we represent them as a sequence of variables that form a discrete Markov model with the transition matrix $T = [p(s_t = l_i|s_{t-1} = l_j]_{i,j=1}^K$. Having labels as unobservable variables and feature vectors as observable variables implies a HMM structure. However, it is difficult to learn a good generative model for the likelihood due to dimensionality of the feature vectors and influences of out-of-the-model variables on feature vectors. Therefore we adopt a hybrid HMM (Fig. 1) which exploits advantages of the discriminatively trained CLLRMP through the likelihood function $p(y_t|s_t) \propto \frac{p(s_t|y_t)}{p(s_t)}$.

3. DYADIC INTERACTION DATASET

The employed dataset is a part of the acted dyadic interaction database [6] that consists of 3 hours of audio, video and motion capture data split in multiple $5-10$min sessions. Each session contains interaction based on unscripted role-play based on one out of 9 conflictual topics that include cheating in relationship, arguing over a drinking problem etc. Topics are selected such that same sex participants act as friends and opposite sex participants as couples. In order to improve chances of recording realistic interactions participants undergo two preparation stages. A couple of days prior to the data collection participants are introduced to the pool of topics and, on the collection day they are asked to discuss topics with their peers and agree on $4-6$ interaction scenarios.

The hardware architecture allows us to record sessions with 10 HD Flea 2 cameras (30fps), 12 sensor Vicon motion capture (MoCap) system (120fps), three 4-microphone T-arrays, two lapel microphones and one shotgun microphone (48kHz) without dropped frames. All modalities are synchronized with a sub 10ms synchronization precision.

Sessions are manually post-processed (to correct errors in MoCap automatic trajectory reconstruction) and annotated in two ways. The first set of annotations include transcription and segmentation of the audio on the speaker turn-taking level augmented by the basic sentence level dialogue acts. The second set of annotations is conducted by trained [1] psychology-domain experts, to provide subject-interaction level labels including, acceptance, presence of blame, attitude and approach-avoidance.

Experts provide us with the continuous-in-time and discrete-in-value approach-avoidance labels for each participant. The approach-avoidance labels belong to an ordered set of nine categories ranging from complete avoidance to the complete approach. Labelers provide two sets of labels, one using only the multi-view video and the other using both video and audio.

Since the labeling and particularly the post-processing are time demanding at present we have 7 fully annotated interaction sessions in total duration of 40 min. This 7-session subset of the full dataset is used in experiments presented in this paper. The AA labels for these 7 sessions belong to a subset $\{-1, 0, 1, 2, 3\}$ of the full label set $\{-4, \ldots, 4\}$.

4. RESULTS AND DISCUSSION

In Section 4.1 we describe details on feature extraction, estimator training and evaluation methodology and in Section 4.2 we present results on different evaluation experiments.

4.1. Features, estimator training and evaluation methodology

For each session and each participant we extract 5 MoCap features: the relative inter-participant head (angle) and body orientation (angle), two measures of the body posture (leaning angle - the angle between spine axis and horizontal plane; body open-closed measure - sum of the triangular square areas defined by elbow, wrist and chest markers for both hands) and the hand velocity measure (maximum of left and right hand velocities). Additionally, we extract two acoustic features, pitch and energy. We get the MoCap features directly from the MoCap marker coordinates every 10 ms and the acoustic features by processing 25 ms speech frames with 10 ms shift. Acoustic features, energy and pitch, are extracted using Praat software. For each feature we get 6 functionals (feature statistics), mean, minimum, maximum, standard deviation, skewness and kurtosis, on 6 s (also 3 s and 4.5 s) functional windows with 1 s shift. Note that the statistics of the audio features can be calculated only in regions where the speaker is active. If a participant does not speak in a particular frame we set all coordinates of the functional vector that correspond to the audio feature statistics to zero. By doing this we avoid occurrences of missing features. For estimation of the video-only and audio-and-video based categories we use, respectively, the 30-dimensional vector of MoCap functionals and the full 42-dimensional functional vector.

Fig. 2. MoCap markers and body/head orientation features.

We split samples for each of 7 sessions in 20 consecutive non-overlapping parts with approximately equal number of samples. We crete 7 train-test set pairs in a leave leave-part-of-session-out manner. Namely, we use 10 odd-indexed parts' of a single session as a test set. Samples from the remaining 10 even-indexed parts are augmented with randomly chosen samples from other sessions to form a balanced training set with 500 samples per AA label. For each train-test set pair we find optimal values for the cost and the width of the radial basis kernel function of SVM classifiers by grid-search using 5-fold cross validation on the training set. We obtain model parameters of the weighted SVM classifiers using weights obtained from the training samples that belong to the same session as the testing samples. Parameter values for the CLLRMP are chosen to be optimal in the maximum likelihood sense. For the SVM training and prediction we used functions from the LibSVM toolkit [10].

4.2. Experiments

First, we present results that demonstrate the influence of the feature processing window length on the average estimation accuracy for different estimation methods (Fig. 3).

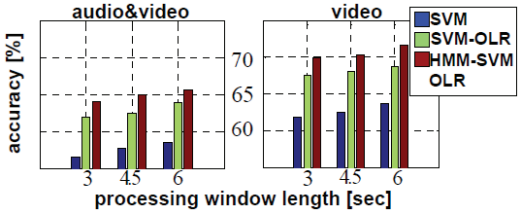

Fig. 3. Dependency of AA category estimation accuracies on feature processing window length for different estimation techniques.

All estimation methods achieve slightly higher accuracies for longer functional windows (Fig. 3). We did not experiment with the windows above 6s for 1s window shift, due to the high correlation between many consecutive feature vectors. The size of the dataset was a limiting factor when considering larger window shifts i.e. less window overlap. The multi-class SVM benefits the most by the increasing window size and HMM-SVM-OLR the least, although the HMM-SVM-OLR accuracy is the highest. This does not come as a surprise, as HMM-SVM-OLR conditions current state on the previous state and therefore exploits context longer than window.

The experts' perception of AA labels differs depending on whether they use video-only or both audio and video in their annotation process. The SVM-OLR suffers 6.1 % lower average accuracy in the estimation of video-and-audio based AA labels when trained on the vision based (MoCap) features than when trained on audio-visual data (MoCap and audio features). This proves that the proposed small set of audio derived non-verbal features captures some of the same information that influences the experts' perception. However, all estimators achieve higher accuracies for the video-only AA labels which may imply that the audio feature set should be extended by additional audio, turn-taking and transcript derived features, but may also imply that there is more variability in interpretations of the audio-based information.

Table 1. Estimation accuracies for AA labels.
(window: 6 s, V and AV:video-only and audio-and-video)

	SVM		**SVM-OLR**		**HMM-SVM-OLR**	
	AV[%]	V[%]	AV[%]	V[%]	AV[%]	V[%]
S_1	60.7	62.3	62.6	69.9	67.2	74.4
S_2	54.1	60.0	60.7	65.3	58.8	65.3
S_3	60.5	63.5	62.6	69.8	65.2	73.1
S_4	57.9	64.8	63.4	70.7	65.5	71.0
S_5	57.4	63.9	60.1	65.1	64.5	69.2
S_6	60.0	65.9	68.2	69.1	69.0	74.1
S_7	57.8	63.9	65.7	69.6	66.0	71.6
AVG	**58.5**	**63.7**	**63.9**	**68.8**	**65.7**	**71.6**

All discussed trends related to difference in the average estimation accuracy for the video-only and the video-and-audio based AA labels remain valid on the session level. Accuracies on all sessions, but S_2 and S_5, are comparable. Although the label prior, obtained from the the part of the testing session that is associated to the training session the training set, represents a good match to the testing label distribution this distribution is much more skewed

than the artificially balanced training set. Namely, its symmetric Kullback-Leibler distance between the AA label distribution P_i for i^{th} testing set and the uniform training AA label distribution P_{-i}, $\frac{1}{2}(KL(P_i\|P_{-i}) + KL(P_{-i}\|P_i))$, takes the smallest values 0.62 and 0.65 when testing on sessions S_2 and S_5, 0.62, 0.65, while its mean and variance for all remaining sessions are respectively 1.36 and 0.22. The indicated mismatch has negative implications on the weighted training for the multi-class SVM and SVM-OLR, and scaling of the HMM-SVM-OLR likelihood.

Further, we analyze importance of the proper treatment of the category ordering and dynamics and its influence on the category confusion patterns. For this purpose we subtract confusion distribution matrices of the SVM baseline for the video-only and the audio-and-video based AA labels from the corresponding matrices for SVM-OLR and HMM-SVM-OLR. Positive diagonal entries of the difference matrices provide insight into class conditioned accuracy improvements and negative off-diagonal elements describe differences in category confusion patterns between proposed methods and the multi-class SVM. We present the difference matrices in form of color maps (see Fig. 4), where dark (light) shades represent negative (positive) values. Similar shades of main diagonal entries

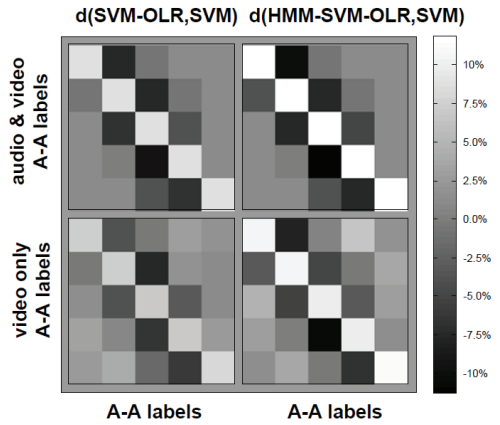

Fig. 4. Differences between confusion matrices (left column: (SVM-OLR) - (SVM); right column: (HMM-SVM-OLR) - (SVM).

in each colormap show that both SVM-OLR and HMM-SVM-OLR improve estimation accuracies for all classes uniformly. Lighter elements on the main diagonal in the right colormap column indicate performance advantage of HMM-SVM-OLR. Dark shade of cells corresponding to the neighboring category pairs, $(c_1, c_2) \in \{(i, j) : |i - j| = \pm 1\}$, show that the CLLRMP is able to distinguish similar categories from the binary SVM outputs.

As explained in Section 2.3 SVM outputs fit the CLLRMP better than the original feature vectors. This is experimentally confirmed by comparison of accuracies for the SVM-OLR and the CLLRMP fitted on the original feature vectors (see Table 4.2).

Table 2. CPL-LR inputs: original features vs. SVM outputs.

	SVM-OLR		**CLLRMP**	
	AV[%]	V[%]	AV[%]	V[%]
AVG	**63.9**	**68.8**	**54.1**	**57.7**

Since the SVM-OLR and the HMM-SVM-OLR exploit label ordering and transitions they improve estimates in situations in which SVM predicts frequent label changes and/or distant label changes in consecutive frames. Therefore, the cells farther from the main diagonal, $(c_1, c_2) \in \{(-1, 1), (0, 2), (2, 0), (3, 1)\}$, get a dark shade.

5. CONCLUSIONS AND FUTURE WORK

We addressed estimation of specific behavioral categories– approach avoidance (AA)– in dyadic human interactions using audio and MoCap derived features. From algorithmic perspective, we proposed two estimation schemes that exploit ordering and dynamics of AA labels, the SVM-OLR and the HMM-SVM-OLR. The SVM-OLR transforms the original feature space by multiple binary SVM classifiers and fits the CLLRMP on classifier outputs. The HMM-SVM-OLR is a hybrid Markov model that uses likelihood function proportional to the ratio of the label posterior probability from SVM-OLR and the label prior. Experimental results on the dyadic interaction dataset show advantages of the ordinal regression methods over the multi-class SVM baseline. Average and single-session estimation accuracies increase for longer feature processing windows. The HMM-SVM-OLR outperforms the SVM-OR and the multi-class SVM and achieves leave-part-of-one-session-out average accuracy of 71.6 % for 6 s window. We discussed: (1) variability in single-session estimation performances; (2) differences between confusion matrices for the proposed estimators and the multi-class SVM; and (3) performance differences when CLLRMP is fitted on SVMs and original feature vectors.

Our ongoing work includes collection and preprocessing of larger datasets of acted couples-therapy and real psychologist-student interactions. Our work in progress focuses on: (a) augmentation of the audio (speech rate, pitch slope and turn taking dynamics) and MoCap (vertical head and body orientation angles) feature sets; (b) inclusion of transcript features and replacement of the analysis oriented MoCap features with features derived from video; (c) analysis of experts' labeling consistency and expert to expert (or non-expert) agreement; and (d) design of time series models that exploit multiple labeler inputs.

6. ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation and the Viterbi Research Innovation Fund.

7. REFERENCES

[1] C. Heavey, D. Gill, and A. Christensen, *Couples interaction rating system 2 (CIRS2)*, University of California, Los Angeles, 2002.

[2] R. E. Heyman, R. L. Weiss, and J. M. Eddy, "Marital interaction coding system: Revision and empirical evaluation," *Behavioural Research and Therapy*, vol. 33, pp. 737–746., 1995.

[3] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, pp. 1743–1759, 2009.

[4] R. Krieglmeyer, R. Deutsch, J. De Houwer, and R. De Raedt, "Being moved: Valence activates approach-avoidance behavior independently of evaluation and approach-avoidance intentions," *Psychological Science*, vol. 21.

[5] R. Adams, N. Ambady, C. Macrae, and R. Kleck, "Emotional expressions forecast approach-avoidance behavior," *Motivation and emotion*, vol. 30, no. 2, pp. 177–186, 2006.

[6] V. Rozgic, A. Xiao, B.and Katsamanis, B. Baucom, P. G. Georgiou, and S. Narayanan, "A new multichannel multimodal dyadic interaction database," in *Proc. IS*, 2010.

[7] A. Agresti, *Analysis of Ordinal Categorical Data*, Wiley, 2010.

[8] E. Frank and M. Hall, "A simple approach to ordinal classification," in *Proc. EMCL*, 2001, pp. 145–156.

[9] P. McCullagh, "Regression models for ordinal data," *Journal of the Royal Statistical Society, Series B*, vol. 42, no. 2, pp. 109–142, 1980.

[10] Chih-Chung Chang and Chih-Jen Lin, *LIBSVM: a library for support vector machines*, 2001, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.