Automatic Identification of Salient Acoustic Instances in Couples' Behavioral Interactions using Diverse Density Support Vector Machines

James Gibson, Athanasios Katsamanis, Matthew P. Black and Shrikanth Narayanan

University of Southern California, Los Angeles, CA, USA

`http://sail.usc.edu`

Abstract

Behavioral coding focuses on deriving higher-level behavioral annotations using observational data of human interactions. Automatically identifying salient events in the observed signal data could lead to a deeper understanding of how specific events in an interaction correspond to the perceived high-level behaviors of the subjects. In this paper, we analyze a corpus of married couples' interactions, in which a number of relevant behaviors, e.g., level of acceptance, were manually coded at the session-level. We propose a multiple instance learning approach called Diverse Density Support Vector Machines, trained with acoustic features, to classify extreme cases of these behaviors, e.g., low acceptance vs. high acceptance. This method has the benefit of identifying salient behavioral events within the interactions, which is demonstrated by comparable classification performance to traditional SVMs while using only a subset of the events from the interactions for classification.

Index Terms: behavioral signal processing, multiple instance learning, diverse density, support vector machines

1. Introduction

In recent years, it has become of interest to use signal processing and machine learning methods to automatically determine humans' affective and behavioral states [1]. There has been an abundance of work devoted to machine understanding of human affect, e.g., emotional state [2], whereas many other facets of human behavior are only beginning to be analyzed in this way. Machine aided analysis of human behavioral interactions could offer a novel tool set to be utilized by diverse domains such as psychology, education, and security. In this paper, we apply a data-driven machine learning approach that uses salient events to classify interactions between married couples enrolled in a therapy study.

In order to assess behavioral interactions, psychologists often rely upon manual coding of audiovisual recordings. Behavioral observations can be time intensive and require a great deal of human effort. In addition to the time and cost limitations of these methods there is also the issue of subjectivity of analysis between evaluators. This inherent subjectivity has been addressed by standard coding manuals such as the Social Support Interaction Rating System (SSIRS) the Couples Interaction Rating System (CIRS) [3, 4], which are used to evaluate married couples' interactions. Although attempts at standardization have been made, agreement between evaluators is still an issue within the field [5].

We propose using signal processing and machine learning techniques to address the aforementioned limitations of human behavioral observation of couple therapy interactions. Automatic classification of couples' therapy interactions using acoustic features was first attempted in [6]. This work demonstrated that acoustic features extracted from couples therapy sessions were suitable for automatically predicting trained evaluators' perceptions of participants' behavior at the session level. These analyses offer a summative session-level assessment of the behavior of interest without offering insights into what aspects of the interaction could have contributed to the resultant judgement.

In addition to classifying behaviors in the interactions at the session-level, it is also of interest to determine salient behavioral events within the sessions. This is of interest because it may offer a more concise representation of the interaction and allow for an analysis of how feature level occurrences correspond to the perceived human behavior. In our work we apply a multiple instance learning (MIL) framework to address this problem. Different formulations of MIL have been applied to many domains including drug activity prediction [7, 8], image categorization [9], and audio classification [10]. We use a particular formulation of MIL known as Diverse Density Support Vector Machines (DD-SVM) that was introduced in [9]. This makes no preconceived designations on what specific acoustic events may be deemed salient, i.e., we do not attempt to detect specific intuitive cues such as laughter or crying. Instead, we utilize the proposed methodology to identify feature level occurrences that are salient to the specific behavioral classification task, without regard to explicit, semantically defined events.

We apply this framework to acoustic features extracted from couples therapy interactions. The experimental results show that this technique achieves comparable classification performance on this dataset to traditional SVMs. Since the DD-SVM framework only uses portions of the interaction deemed to be salient, this comparable performance implies that the DD-SVM method is able to detect regions that are most relevant for the classification of the behavioral codes.

2. Corpus

The corpus was collected as part of a joint longitudinal study on couples therapy between the University of California, Los Angeles and the University of Washington [11]. The study included recording 134 chronically distressed married couples at various intervals over a one year therapy period: before therapy, 26 weeks into therapy, and two years after therapy. The corpus consists of 569 ten-minute dyadic interactions (between husband and wife), in which the married couples discussed a problem in their relationship.

Each session was recorded with a single far-field microphone (16 kHz, 16-bit). Because the recordings were not originally intended for automatic analysis, audio specifications (e.g., microphone placement, environmental conditions) were highly variable across sessions. For this reason, sessions with an average signal-to-noise ratio (SNR) less than 5 dB were omitted from this analysis. The sessions were transcribed with the speaker explicitly labeled (wife or husband). No timing information was marked in the transcriptions. A more detailed description of the corpus can be found in [6].

As part of the original study, the sessions were coded according to 33 behavioral codes: 20 codes from the SSIRS and and 13 from the CIRS. Both coding systems rated individual spouses's behavior at the session-level on a 1-9 scale (1 corresponding to low occurrence of a particular behavior and 9 corresponding to high occurrence). Each session was manually coded by three to four trained evaluators. The six codes with the highest inter-evaluator agreement were then chosen for automatic classification: level of acceptance toward the other spouse, level of blame, global positive affect, global negative affect, level of sadness, and use of humor.

One necessary preprocessing step taken before we could extract meaningful acoustic features for each spouse was segmenting the sessions into individual speaker regions. Rather than manually segmenting the corpus, we exploited the existence of the transcriptions (with speaker labels) and used a recursive speech-text alignment procedure implemented in freely-available software we developed, *SailAlign* [12]. After convergence, the session was split into wife regions, husband regions, and unknown regions in which we were unable to align the audio to the transcription. For this paper, we ignored all sessions in which we could not segment at least 55% of both the wife's and husband's transcribed words into single speaker regions. After taking into account the 5 dB and 55% speaker segmentation thresholds, 372 sessions remained for analysis (65% of the original corpus).

In accordance with previous and ongoing work with this corpus, we framed the problem of behavioral rating prediction as a binary classification task. For each code and gender pair separately we identified the sessions in which the corresponding spouse had mean code scores (averaging across evaluators) that fell in the top 70 and bottom 70 of the score range (approximately the top and bottom 20%) and selected those for training and testing our classifiers. This split produced 12 gender-and-code dependent subsets, i.e., 6 for the husbands and 6 for the wives, of 140 sessions, each comprising 70 "high" examples of the code and 70 "low" examples for the specific code-gender pair.

3. Methodology

At an abstract level, each of the recorded sessions to be analyzed corresponds to a sequence of behavioral manifestations. For example, one of the spouses may initially be exhibiting a low level of acceptance towards the other spouse but this may gradually change as the interaction unfolds. By the end of the interaction, this change could be great enough to result in the spouse being rated as highly accepting at the session-level despite the fact that the corresponding behavior is not apparent throughout the whole session. Clearly, automatic prediction of the session-level code, i.e., classification into either high or low level of acceptance, should exploit the fact that there are certain instances in the interaction where the behavior is more strongly displayed. These instances are considered to be the most salient for the particular task but typically are not specifically annotated as such. What we have is just the session-level code and, for that reason, an instance-level salience model cannot be developed in a supervised manner. This problem falls into a general category of problems that is commonly referred to as multiple instance learning. Our goal is to identify the instances of the interaction that make the largest impact and rely on them for the behavioral classification of the entire interaction.

Adopting the relevant machine learning terminology we refer to each session $\mathbf{B}_i$ as a bag and we assume that it comprises several behavioral instances, i.e., $\mathbf{B}_i = \{B_{i1}, B_{i2}, ...B_{iN_i}\}, \forall i = 1, 2, ..., L$, where N_i is the number of instances in bag i and L is the number of training bags. Each behavioral instance, B_{ij}, is represented by a $N_f \times 1$ vector of features. In our work, we assume that each instance corresponds to a speaking turn of the spouse being evaluated. Each bag has a label, $+1$ or -1 for high or low rating respectively of the behavioral code of interest. The set of labels is denoted as $\mathbf{Y} = \{y_1, y_2, ..., y_L\} : y_i \in \{+1, -1\} \forall i$. Given the specificities of the problem, we apply Diverse Density Support Vector Machines to predict the bag labels from the instance observations.

3.1. Diverse Density Support Vector Machines

For each spouse and behavioral code combination we use the DD-SVM algorithm to classify the session-level label of the spouse being assessed. The DD-SVM method is briefly described below. A more in-depth description of the algorithm can be found in [9].

1. **Estimate the diverse density for each instance.** We want to identify regions where there is a high concentration of instances from different bags of the same label and which are far from instances from bags of the opposite label. For this purpose, we first estimate the diverse density at each instance which is defined as [13, 9]:

$$DD(\mathbf{x}, \mathbf{w}) = \prod_{i=1}^{L} \left[\frac{1 + y_i}{2} - y_i \prod_{j=1}^{N_i} \left(1 - e^{-||B_{ij} - \mathbf{x}||^2_{\mathbf{w}}} \right) \right], \tag{1}$$

where $||\mathbf{x}||_{\mathbf{w}} = [\mathbf{x}^T \text{diag}(\mathbf{w})^2 \mathbf{x}]^{\frac{1}{2}}$. In this equation, $\mathbf{x}$ is the feature vector representing an instance prototype and $\mathbf{w}$ is a weight vector to compensate for the fact that the features in $\mathbf{x}$ may not be equally important.

2. **Maximize the diverse density.** We then find local maximizers to the diverse density using the Expectation Maximization-Diverse Density (EM-DD) algorithm, as originally introduced in [8].

3. **Prototype selection.** After the diverse density is maximized, instance prototypes of insufficiently high diverse density are rejected. Then we select local maxima from each region of high diverse density to represent that region. These representative instance prototypes are then regarded as the "salient" prototypes from these regions. The size of each region is controlled by an independent parameter β which has to be chosen appropriately [9].

4. **Bag feature computation and classification.** Once selected, the N_p salient instance prototypes and their corresponding weights, denoted by starred pairs $(\mathbf{x}_k^*, \mathbf{w}_k^*), k = 1, 2, ..., N_p$, are used to compute the features to represent each bag. The feature vector for a given bag/session, $\phi(\mathbf{B}_i)$, defined as:

$$\phi(\mathbf{B_i}) = \begin{bmatrix} \min_{j=1,\dots,N_i} ||B_{ij} - \mathbf{x}_1^*||_{\mathbf{w}_1^*} \\ \min_{j=1,\dots,N_i} ||B_{ij} - \mathbf{x}_2^*||_{\mathbf{w}_2^*} \\ \vdots \\ \min_{j=1,\dots,N_i} ||B_{ij} - \mathbf{x}_{N_p}^*||_{\mathbf{w}_{N_p}^*} \end{bmatrix}, \tag{2}$$

and corresponds to the minimum weighted distance of the bag, to each of the salient instance prototypes. This results in an $N_p \times 1$ feature vector representation of each session, where N_p is the number of salient prototypes. Using these features, we classify the sessions in a standard SVM-based framework.

3.2. Acoustic Features

For this study we chose a much smaller feature set than the one employed in [6]. This is because the aim of our current study is to investigate the selection of salient regions by applying the DD-SVM technique and not to determine the most complete acoustic representation. Mel-Frequency Cepstral Coefficients (MFCCs) were chosen because they can be robustly extracted and they have been shown to perform well in the behavioral classification task.

In our setup, 14 MFCCs were extracted every 10 ms over 25 ms Hamming-windowed frames using the openSMILE software [14]. The MFCCs were computed using a bank of 26 triangular filters evenly centered on the mel-scale from 20 to 8000 Hz. Once extracted, the MFCCs are separately mean-normalized for each speaker in every session. The feature vector B_{ij} representing the speaker turn/instance j in the session/bag i comprises the means and variances of the MFCCs in the particular turn. These 28-dimensional feature representations of behavioral instances are then used to select the instance prototypes and train the SVM as described in Sec. 3.1.

4. Experimental Results

Ten fold cross validation across the couples was used in all our classification experiments. So, sessions from the same couple could only appear either in the training or in the testing set, but not in both. As a baseline, we chose to use a conventional SVM method where each session B_i is represented by an extended feature vector $\phi_{ext}(B_i)$ that includes the minimum distances to all possible instances in the training set and not only to those selected using the diverse density maximization process. As in [9] the Gaussian kernel, $K(a, b) = e^{-\gamma ||a-b||^2}$, is used. The parameters γ, C for both the baseline and the diverse density SVMs, as well as the parameter β used in the prototype selection process for the latter one are chosen using two fold cross validation on the training set.

Table 1 shows the average classification accuracy across the ten folds. The average classification accuracy, across the behavioral code-spouse configurations, is 67.9% (± 7.9%) using DD-SVMs versus 66.9% (± 8.4%) using traditional SVMs. The percentage of instance prototypes selected using the diverse density based process is shown in Table 2. On average, across the behavioral code-spouse pairs, 26.8% of the prototypes were kept for classification. This is a large reduction in the amount of data provided for the classification task with an average increase in classification performance.

To further elaborate on the diverse-density based instance prototype selection process, in Fig. 1 we provide scatter-plots of the training sessions for a randomly chosen fold for the code humor. For visualization purposes, each session B_i is represented by only two components of the vector $\phi(B_i)$. The “high humor” rated sessions are marked with a ‘+’ , while the “low humor” ones are marked with a ‘○’. In Fig. 1(a), the components corresponding to the most salient positive instance prototype and most salient negative instance prototype were chosen, i.e., to those with the maximum diverse density. It is evident that although these prototypes alone do not provide clear discrimination between the regions, they do offer some apparent separation of the sessions from different class. The components chosen in Fig. 1(b) for the representation correspond to the instance prototypes with the minimum diverse density. There is no visible discrimination between regions demonstrating that session features computed based on these instance prototypes would possibly deteriorate overall classification performance. Figure 1(c) gives the corresponding trainining set representation when two prototypes with approximately the average diverse density of all instance prototypes are used for the reduced $\phi(B_i)$ estimation. While there is significantly more discrimination between classes than in Fig. 1(b), it is not to the extent displayed in Fig. 1(a). Sessions in Fig. 1(d) are represented by distance vectors from two randomly selected prototypes. Overall, these figures support the validity of using the diverse density to systematically select salient instance prototypes for our classification task.

Table 1: *Average classification accuracy (%) of behavioral codes.*

Configuration	Wife		Husband	
	DD-SVM	SVM	DD-SVM	SVM
acceptance	**73.6**	72.1	69.3	69.3
blame	77.1	**79.3**	**72.3**	71.4
positive	**74.3**	70.0	55.0	**58.6**
negative	75.0	**77.9**	**71.4**	70.0
sadness	**66.4**	57.1	**63.6**	62.9
humor	**52.9**	51.4	63.6	63.6

Table 2: *Percentage of instances used for DD-SVM classification (average across folds). The baseline SVM classification used 100% of the instances.*

Configuration	Wife	Husband
acceptance	40.5	26.8
blame	41.7	35.4
positive	28.7	43.4
negative	34.9	18.9
sadness	6.0	18.5
humor	4.1	22.3

5. Conclusions and Future Work

In this work, we demonstrated that the DD-SVM algorithm offers a promising framework for the couples’ behavioral interaction classification task. This suggests that expert evaluators’ perceptions of certain human behaviors can be accurately modeled when only instances within an interaction that are salient to that behavior are considered. In Table 2, we show a comparison of the percentage of prototypes that are kept as salient for each code/spouse type configuration. It is apparent that different behavioral codes require significantly different percentages of the total number of instances for classification. For example, when classifying wives’ level of acceptance 40.5% of the instances are kept as salient whereas only 6.0% are kept for classifying wife’s level of sadness. This suggests that certain perceived human behaviors, as represented by acoustic features derived from speech, can be modeled by only regarding representative examples of that behavior.

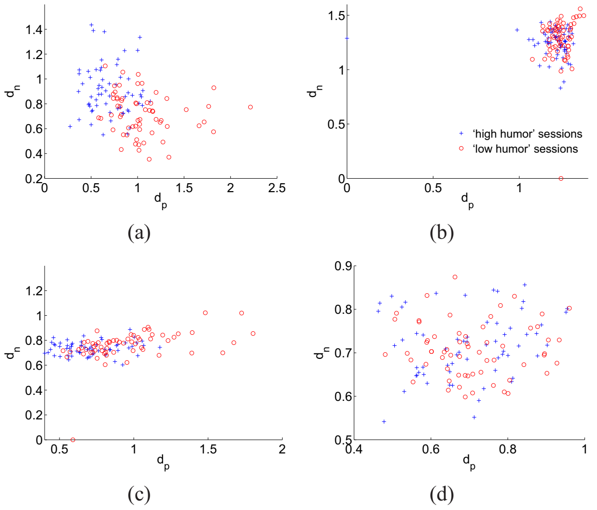

Figure 1: *Two-dimensional training set representation based on a reduced feature vector for each session for the wives' level of humor configuration. The axes, d_p and d_n, represent the distance of the training bags from the positive and negative instance prototype of selected diverse density.*

It should be noted that the proposed method did not yield improved classification performance in all behavioral code, spouse configurations. In these cases (wife's level of blame, wife's global negative affect, and husband's global positive affect) the difference in accuracy is small (2.2%, 2.9%, and 3.6%, respectively) compared to the reduction in number of instances used for classification (52.9%, 65.1%, and 56.6%). This is most likely due to the automatic procedure by which instance prototypes are determined to be non-salient. As part of our future intended work, we hope to improve the tuning of this procedure to ensure that no salient instances are disregarded.

It is also important to note that the instances selected as salient prototypes are done so with regard to the signal-derived features used for classification, in this case MFCCs. The prototypes with maximum diverse density may change to some degree depending on how representative a certain feature set is of the underlying behaviors taking place in these interactions. Therefore in the future, we plan to explore different feature sets and how they affect which prototypes are selected as salient to the classification task.

In the future we would like to further study the aspect of saliency in human behavioral interactions. This can be accomplished by introducing more feature sets, exploring different approaches/frameworks, and altering our underlying definitions of what constitutes a behavioral instance. With regard to feature sets, in [6] we used a very large feature set ($N_f = 2007$) and demonstrated the merit of various feature combinations. While this approach may not be tractable in our current configuration, we plan to try other feature subsets that performed well. It is also of interest to choose a feature set that will lend itself to some higher level interpretability. Initial plans will incorporate prosodic features such as pitch and energy.

Our fundamental definition of a behavioral instance as a speaker turn within an interaction may be revised in future studies. For example, we intend to investigate how salient events compare when the duration of a behavioral instance is defined by computational methods such as a sliding time window of constant or varying width or by linguistically informed events such as spoken words or syllables. We are also interested in exploiting a priori information in saliency detection. However, this approach will first require a data set where interactions are coded at a finer level of granularity than the session-level.

Finally, we would like to fuse this approach with different modalities such as session transcripts and visual data. A multimodal approach may offer a fuller representation of behavioral interactions, just as humans utilize multiple modalities to understand and interpret one another's behavior.

6. Acknowledgements

We would like to thank Brian Baucom and Andrew Christensen for giving us access to the couple therapy dataset. This research is partially supported by the National Science Foundation and the USC Annenberg Graduate Fellowship Program.

7. References

[1] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[2] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. Interspeech*, 2009, pp. 312–315.

[3] J. Jones and A. Christensen, *Couples interaction study: Social support interaction rating system*, University of California, Los Angeles, 1988.

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

[5] G. Margolin, P. Oliver, E. Gordis, H. O'Hearn, A. Medina, C. Ghosh, and L. Morland, "The nuts and bolts of behavioral observation of marital and family interaction," *Clinical Child and Family Psychology Review*, vol. 1, no. 4, pp. 195–213, 1998.

[6] M. Black, A. Katsamanis, C. Lee, A. Lammert, B. Baucom, A. Christensen, P. Georgiou, and S. Narayanan, "Automatic classification of married couples' behavior using audio features," in *Proc. Interspeech*, 2010.

[7] T. Dietterich, R. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[8] Q. Zhang and S. Goldman, "Em-dd: An improved multiple-instance learning technique," *Advances in neural information processing systems*, vol. 2, pp. 1073–1080, 2002.

[9] Y. Chen and J. Wang, "Image categorization by learning and reasoning with regions," *The Journal of Machine Learning Research*, vol. 5, pp. 913–939, 2004.

[10] K. Lee, D. Ellis, and A. Loui, "Detecting local semantic concepts in environmental sounds using markov model based clustering," in *Proc. ICASSP*, 2010, pp. 2278–2281.

[11] A. Christensen, D. Atkins, S. Berns, J. Wheeler, D. Baucom, and L. Simpson, "Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples," *Journal of Consulting and Clinical Psychology*, vol. 72, no. 2, pp. 176–191, 2004.

[12] A. Katsamanis, M. Black, P. Georgiou, L. Goldstein, and S. Narayanan, "Sailalign: Robust long speech-text alignment," in *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, 2011.

[13] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Advances in neural information processing systems*. Citeseer, 1998, pp. 570–576.

[14] F. Eyben, M. Wollmer, and B. Schuller, "OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit," in *3rd International Conference on Affective Computing and Intelligent Interaction*, 2009, pp. 1–6.