Inferring truth from multiple annotators for social interaction analysis

Gokul Chittaranjan
Idiap Research Institute and EPF Lausanne
gthatta@idiap.ch

Oya Aran
Idiap Research Institute
oaran@idiap.ch

Daniel Gatica-Perez
Idiap Research Institute and EPF Lausanne
gatica@idiap.ch

Abstract

This study focuses on incorporating knowledge from multiple annotators into a machine-learning framework for detecting psychological traits using multimodal data. We present a model that is designed to exploit the judgements of multiple annotators on a social trait labeling task. Our two-stage model first estimates a ground truth by modeling the annotators using both the annotations and annotators' self-reported confidences. In the second stage, we train a classifier using the estimated ground truth as labels. Our experiments on a dominance estimation task in a group interaction scenario on the DOME corpus, in addition to synthetically generated data, give satisfactory results, outperforming the commonly used majority voting as well as other approaches in the literature.

1 Introduction

In many studies conducted in the areas of affective and social behavior, researchers are interested in finding psychological traits. Traits are defined as habitual patterns of behavior, thought and emotion [8] that are often identifiable in group interactions. In the recent past, there have been many studies on automated methods for detecting important traits in face-to-face group interactions such as dominance [11, 1, 7] and other traits related to personality [9]. In these studies, multiple annotators are often asked to label the traits. The availability of online annotation resources (e.g. Mechanical Turk) opens up the possibility of obtaining multiple human judgments for each data point. These judgments have to be handled with care [3]. In the standard psychology literature, the quality of these annotations is determined via inter-rater agreements, often measured using metrics such as Cohen's Kappa values [4]. If the inter-rater agreement is sufficiently high, a common approach is to use the majority agreement of annotators as the ground-truth labels. However, the removal of no-agreement cases leads to the shrinkage of data sets that are already moderately sized. In majority voting, all annotators are weighted equally, whereas in reality, some might do better than the others or may express more confidence than others. The main challenge in estimating the "true" label from these annotations is that the expertise of the annotators is not readily quantifiable. Moreover, given that they are all human judgments, the "perfect" ground-truth might not be available for even a subset of the data, making validation of the estimated labels impossible.

In this paper, we introduce models that circumvent these problems and have the potential of being applied to many instances of affective and social interaction analysis. We use the knowledge provided by the annotators, the annotations and their confidences, to estimate final class labels and then use these to train classifiers. For experimental verification, we apply it to the case of identifying dominant persons in small group conversations [7] from nonverbal communicative cues extracted from audio and video. To show the importance of using annotator confidences in modeling annotators, we benchmark our model with the presence and absence of this information on a recent, publicly available corpus of group interactions (DOME) [2]. Finally, we compare it to the integrated single-step model introduced by Raykar et. al [10]. Although we discuss the results with dominance data, the models are easily generalizable to other domains that use multiple annotations. An extended version of this paper is currently under evaluation at "Automatic Face and Gesture Recognition 2011".

2 Related Work

Rienks and Heylen [11] were among the first to study dominance in group conversations using computational means. More recently, Jayagopi et.al. [7] and Aran and Gatica-Perez [1] introduced a set of nonverbal features that could be used for detecting the most dominant person in meetings on a subset of the popular Augmented Multi-Party Interaction (AMI) corpus [5], for which multiple observers provided dominance judgement. These studies had several shortcomings. Only meetings with majority and full agreement annotations were used and no analysis was done to detect whether there were more than one dominant person in the meeting.

While researchers working on computational tasks with subjective annotations have mainly used the Kappa statistic as a measure of annotation reliability [4], and estimated the class labels via majority voting, some studies have attempted to model these annotations to estimate the underlying "true" label. Smyth et al. [12] proposed an EM based model to estimate the reliability of the annotators on an image labeling task. Raykar et. al. [10] recently introduced a method for modeling annotators to obtain estimates of the true class label while jointly modeling a classifier.

To our knowledge, no such methods have been applied to social interaction data. Additionally, the single-step approach of jointly modeling the classifier and the annotations [10] does not incorporate the confidence expressed by the annotators. In the area of human interaction analysis, one of the objectives is to find new features that can be used for further analysis. This means that the introduction of a new feature requires re-modeling the annotators, since there is a close coupling between the input features and the annotator knowledge in this model. Our model allows to include annotation weights into the framework to explicitly model the dependency of the annotations on the "difficulty" of the task. It also allows to train classifiers that use features in a separate step, enabling us to compare the results of classifier with a "derived ground-truth". Moreover, it gives the flexibility of using any classification method, without a need to directly connect it to the annotator model.

3 The Model

For our task, we use a two-stage model (hereafter referred to as Model I). The block diagram of this model is given in Fig. 1. The first stage involves the modeling of annotators (hereafter called the A-model). The A-model is used to obtain the class label estimates, that can be used as ground truth for further analysis. We consider two kinds of information in the annotations: (i) The label indicating the choice of the annotator for the most dominant person (ie. the person having rank 1), with other participants (having lower ranks) in the meeting being labeled as not dominant, and (ii) the dominance weight for the participants in the meeting. This model is inspired from the work of Raykar et. al. [10] with the difference being that the features used to train the joint classifiers are replaced by the dominance weights given to a participant by the annotators. If we have N samples and R annotators, each annotator j is modeled by two parameters called the sensitivity (α_j) and specificity (β_j), which are defined as:

$$\alpha_j = \mathrm{Pr}[y_i^j = 0 | z_i = 0] \tag{1}$$

$$\beta_j = \mathrm{Pr}[y_i^j = 1 | z_i = 1] \tag{2}$$

where z_i is the true label for the sample i and y_i^j is the annotation given by annotator j for sample i. Given the annotations and the dominance weights ($\mathcal{D} = [\{y_i^j | i \in [1, N], j \in [1, R]\}, \{x_i | i \in [1, N]\}]$, to obtain these parameters and consequently, estimates of the ground truth, we carry out the following optimization:

$$\theta_{\mathbb{ML}} = \arg\max_{\theta} Pr[\mathcal{D}|\theta] \tag{3}$$

These estimates can be obtained using a combination of EM and Newton-Raphson update [10].

The A-model estimates are then used in the second stage of our model, in order to train a classifier in a supervised manner, using audio-visual features (the F-model). For our experiments, we have

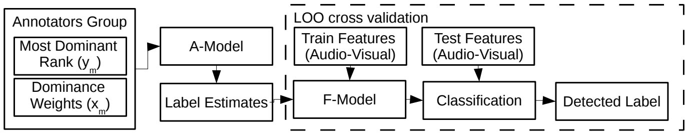
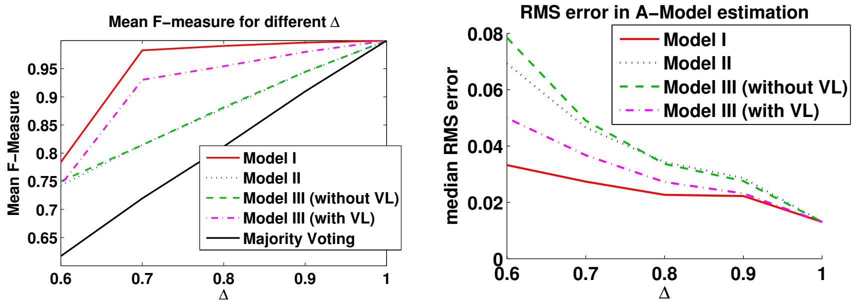

Annotators Group
Most Dominant Rank (y_m)
Dominance Weights (x_m)
A-Model
Label Estimates
LOO cross validation
Train Features (Audio-Visual)
Test Features (Audio-Visual)
F-Model
Classification
Detected Label

Figure 1: The proposed two-stage model



(a) (b)

Figure 2: Mean F-measures for estimating true labels with A-models for synthetic data. (a) Mean F-measure averaged across γ_3, given Δ. (b) Median RMS error in model estimation for different Δ

thresholded the labels into binary class labels and then used a Support Vector Machine (SVM) with radial basis kernel to perform this function.

In order to emphasize the importance of the use of annotation weights, we compare the estimated labels by the A-Model with the method proposed in [12] (Model II). Finally, we compare the results with the integrated one-stage model introduced in [10] (Model III).

4 Experimental Setup and Results

We performed our experiments on a synthetic dataset followed by the DOME corpus for dominance estimation. The synthetic dataset was used for validating our model's performance with respect to a known ground truth, as ground truth does not exist for the dominance estimation task. On the DOME corpus, we extracted several audio, visual and joint audio-visual features as descriptors of dominance, as introduced by Jayagopi et al. [7] and, Aran and Gatica-Perez [1]. These features are defined based on social psychology findings, stating that dominance is often displayed via audio and visual cues such as speaking time, turns, interruptions, visual activity, expressions and gaze [6].

4.1 Results on Synthetic Data

To generate the synthetic data, we defined a new task on the DOME corpus: "Determine whether the participant is visually more active than the average group activity' by labeling particpants having a visual activity length (VL) of more than 25% of the total activity in a meeting as "visually active". The weights given by these annotators were computed by adding noise to the ground truth weights.

To study the effect of different values of α and β, we varied these parameters and calculated the mean F-Measure of the estimated labels with respect to the ground truth. For simplicity, we assumed equal values of sensitivity and selectivity ($\alpha_j = \beta_j = \gamma_j$). Also two annotators were fixed to have the same γ ($\gamma_1 = \gamma_2 = \Delta$) and γ_3 was varied. The results for all our experiments were based on the mean value obtained for 10 independent trials. It is clear from Fig. 2(a) that Model I outperforms Models II and III. We see that the exclusion of VL when training Model III causes the label estimation accuracy to drop below that of Model II, but its inclusion improves it beyond that of Model II. This is an expected result as the ground truth has been derived from the VL feature.

Upon plotting the RMS error in the A-model parameter estimation averaged across all γ_3 values, for a given value of Δ (given in Fig. 2(b), we see that Model I outperforms both Models II and III. The errors converge for $\Delta = \gamma_3 = 1$, which is expected as the annotations are very clean in this case and the additional information given in the form of annotation weights or other features is not required. The RMS errors for Model III trained without VL is higher than Model III trained with VL. An important implication of this result is that an integrated model as suggested by Raykar et.al. [10] works well when the features are not noisy. This further justifies the use of our two stage model for the next experiment on the DOME corpus with real annotators.

Table 1: F-measure performance of F-model on audio-visual features

Model	Feature Set					
	Audio (A)	Video (V)	Audio-Visual (AV)	A+V+AV	A + V	Baseline*
Model I	0.792	0.724	0.803	0.521	0.779	0.395
Model II	0.770	0.726	0.771	0.450	0.680	0.411

* Case when majority class is chosen always

4.2 Results on the DOME corpus

The DOME corpus is composed of 125 four-person meeting segments, that were annotated by 14 annotator groups of size three. On average, about 26 meetings were annotated by each group. For our original model, we trained one A-model per annotator group in the first stage. In the second stage, we used all the estimated labels together to train the classifier that uses audio-visual cues described by Aran and Gatica-Perez [1]. Since our data set size was moderate ($125 \times 4 = 500$ data samples), we used leave one out cross validation to obtain average performance. A threshold of 0.5 was used to obtain binary class labels from the probabilistic scores. SVMs were then trained with these binary labels (results in Table 1. We chose mean F-measure as our performance metric because of the imbalance in the class sizes, as accuracy does not take this into account. We found that Model I consistently performed better than Model II for all features except visual features, in which case, both gave nearly the same performance, justifying the use of annotator weights in the model.

5 Conclusion

In the context of social interaction analysis, we have addressed the open problem of modeling annotators' knowledge to obtain estimates of the ground truth and then training a classifier. This enables us to use the "derived ground truth" for modeling the data using audio/visual features separately. Further, annotation weights, as used here, were found to be quite useful. By using our approach, all the annotated data can be used, without the need for majority agreement. Finally, our method removes the restriction that there could be only one person showing dominance, which was a shortcoming in past work [1, 7, 2]. Our best model used audio-visual (AV) cues, with a performance measure well above the baseline. In the future, we plan to study the relationship between these models and inter-rater agreement values. The generalization of this method to other forms of social data requires further investigation.

Acknowledgments

This work was supported by EU FP7 MC IEF project NOVICOM and Swiss NSF project SONVB.

References

[1] O. Aran and D. Gatica-Perez, "Fusing audio-visual nonverbal cues to detect dominant people in small group conversations," in *ICPR*, 2010.

[2] O. Aran, H. Hung, and D. Gatica-Perez, "A multimodal corpus for studying dominance in small group conversations," in *LREC MMC*, 2010.

[3] M. D. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? (pre-print)," *Perspectives on Psychological Science*, In press.

[4] J. Carletta, "Assessing agreement on classification tasks: The Kappa statistic," *Computational Linguistics*, pp. 249–254, 1996.

[5] J. Carletta et al., "The AMI meeting corpus: A pre-announcement," in *MLMI*, 2005.

[6] J. Hall, E. Coats, and L. LeBeau, "Nonverbal behavior and the vertical dimension of social relations: A meta analysis." *Psychological Bulletin*, p. 1313(6):898–924, 2005.

[7] D. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, "Modeling dominance in group conversations using nonverbal activity cues," *IEEE Transactions on Audio, Speech and Language Processing*, 2009.

[8] S. Kassin, *Psychology*. Prentice-Hall, Inc., 2003.

[9] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro, "Multimodal recognition of personality traits in social interactions," in *ICMI*, 2008.

[10] V. Raykar, S. Yu, L. Zhao, G. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *Journal of Machine Learning Research*, 2010.

[11] R. J. Rienks and D. Heylen, "Automatic dominance detection in meetings using easily detectable features," in *MLMI*, 2005.

[12] P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi, "Inferring ground truth from subjective labelling of venus images," in *NIPS*, 1995.