INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

An Expectation Maximization approach to Joint Modeling of Multidimensional Ratings derived from Multiple Annotators

Anil Ramakrishna[1], *Rahul Gupta*[1], *Ruth B. Grossman*[2], *Shrikanth S. Narayanan*[1]

[1]Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA, USA
[2]Department of Communication Sciences and Disorders, Emerson College, Boston, MA, USA
akramakr@usc.edu, guptarah@usc.edu, ruth_grossman@emerson.edu, shri@sipi.usc.edu

Abstract

Ratings from multiple human annotators are often pooled in applications where the ground truth is hidden. Examples include annotating perceived emotions and assessing quality metrics for speech and image. These ratings are not restricted to a single dimension and can be multidimensional. In this paper, we propose an Expectation-Maximization based algorithm to model such ratings. Our model assumes that there exists a latent multidimensional ground truth that can be determined from the observation features and that the ratings provided by the annotators are noisy versions of the ground truth. We test our model on a study conducted on children with autism to predict a four dimensional rating of expressivity, naturalness, pronunciation goodness and engagement. Our goal in this application is to reliably predict the individual annotator ratings which can be used to address issues of cognitive load on the annotators as well as the rating cost. We initially train a baseline directly predicting annotator ratings from the features and compare it to our model under three different settings assuming: (i) each entry in the multidimensional rating is independent of others, (ii) a joint distribution among rating dimensions exists, (iii) a partial set of ratings to predict the remaining entries is available.
Index Terms: Multiple Annotator Modeling, Expectation Maximization algorithm.

1. Introduction

In several machine learning domains including speech and spoken language based applications, obtaining labeled data attributes can be very expensive and/or cumbersome while unlabeled data points are usually available in abundance. This constrains the direct application of several traditional supervised learning techniques and calls for specialized methods. A few popular techniques that address this problem include active learning [1], domain adaptation [2] and crowd sourcing [3]. In particular, crowd sourcing focuses on pooling ratings from naive annotators and combining them to estimate the unknown ratings. However, the combination is often ad-hoc (eg.: use of majority voting, mean of annotator ratings) and ignores valuable annotator specific information. Previous works have addressed this issue and presented frameworks for modeling the annotators' behavior. In particular, Dawid [4] proposed a model assuming the ground truth to be a latent variable and the annotator judgments to be noisy functions of the latent ground truth. Raykar et al. [3] extended this model to a discriminative case incorporating dependency of the ground truth of an instance to a set of features corresponding to that instance and several works [5, 6] have made further additions to this model. However, explorations to the case when annotators provide a multidimensional rating have not been made. In this paper, we propose an extension of this model to the case of multidimensional ratings. Through our model, we aim to exploit the dependency between not only the annotators but also the entries in the multidimensional ratings. The goal of this work is an accurate prediction of the ground truth to address the issues of rating cost and cognitive load on the annotators. We test several settings of our model incorporating various independence assumptions and partial availability of a few rating dimensions and present our results.

Several previous works have modeled ratings from multiple annotators, each assuming a different dependency structure between the ratings for a data instance and the ground truth corresponding to it. In the model proposed by Raykar et al. [3], a discriminative function encodes the relation between a data instance's features and the ground truth. Furthermore, ratings from each annotator are considered to be noisy versions of the ground truth. Audhkhasi et al. [6] extended this model considering the distribution of the annotator noises to be dependent on the distribution of the features in the data. Gupta et al. [5] proposed an extension of model proposed by Raykar et al. to model continuous time-series ratings from multiple annotators. On the other hand modeling multidimensional ratings has been well studied in the field of multitask learning [7]. Example applications include natural language processing [8], medical risk evaluation [9] and phoneme recognition [10]. Despite the independent progress in multiple annotator modeling and multitask learning, the problem of modeling multidimensional ratings from multiple annotators has not been investigated in the past. We attend to this issue as the subject of this study.

Akin to the training algorithm suggested by Raykar et al. [3], our model is also trained using an Expectation Maximization (EM) algorithm [11]. The EM algorithm is an iterative procedure in which we first estimate the hidden ground truth for the multidimensional ratings (E-step) and use it to compute the model parameters (M-step). We test the model on the Safari Bob dataset [12], which involves children watching and imitating emotional expressions from a video. The videos are annotated by multiple annotators over Amazon Mechanical Turk (M-Turk) on the dimensions of expressivity, naturalness, pronunciation goodness and engagement. To evaluate the model, we make predictions on the annotator ratings over the four dimensions. We further present an extension to our model using which we can reduce the number of entries queried in the multidimensional rating to each annotator. This is desirable both in terms of cost effectiveness as well as reducing the cognitive load on the annotators. We test our model for accurate annotator prediction under three different settings of the proposed model: (i) joint annotator-independent rating modeling (ii) joint annotator - joint rating modeling and, (iii) conditional modeling assuming partial availability of a few dimensions from the multidimensional rating. We compare these three settings to a baseline model directly modeling each annotator individually. Using these three modeling schemes we aim to answer questions related to the improvements obtained from collective modeling of attributes over independent modeling and, improvements in prediction with the availability of judgments on a subset of attributes on the remaining attributes. We show that the models we propose perform better than the chosen baseline. We also perform a follow up experiment by subsequent removal of annotators with fewer ratings and comment on the gains obtained for each model setting.

Copyright © 2016 ISCA

http://dx.doi.org/10.21437/Interspeech.2016-270

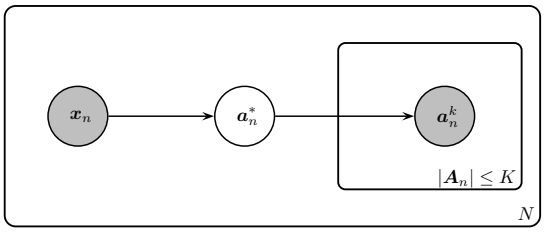

Figure 1: Graphical model representation for the proposed model. $\boldsymbol{x}_n$ is the set of features for the n^{th} instance, $\boldsymbol{a}_n^*$ is the latent ground truth and $\boldsymbol{a}_n^k$ is the rating provided by the k^{th} annotator for that instance. $\boldsymbol{x}_n$ and $\boldsymbol{a}_n^k$ are observed variables, $\boldsymbol{a}_n^*$ is latent. $\boldsymbol{A}_n$ is the set of annotator ratings for the n^{th} instance.

2. Multiple annotator modeling

Consider a set of N data points with features $\{\boldsymbol{x}_1, .., \boldsymbol{x}_N\}$; $\boldsymbol{x}_n$ being the feature vector corresponding to the n^{th} instance. Each data point is associated with a D dimensional ground truth for which judgment from several annotators are pooled. In this work, we assume that each datapoint is annotated by a subset of K annotators. This is a more general setting than assuming that the ratings are available from every annotator (as assumed in [3]), and is often the case with data collection over online platforms such as M-Turk. We represent the set of ratings for the n^{th} data point by a set $\boldsymbol{A}_n$. For example, if annotators 1,2 and 5 provided their ratings (out of K annotators), $\boldsymbol{A}_n$ would be the set $\{\boldsymbol{a}_n^1, \boldsymbol{a}_n^2, \boldsymbol{a}_n^5\}$, where $\boldsymbol{a}_n^k$ is the multidimensional rating from the k^{th} annotator. The vector $\boldsymbol{a}_n^k$ is a D-dimensional vector, represented as $\{a_{n,1}^k, .., a_{n,d}^k, .., a_{n,D}^k\}$, where $a_{n,d}^k$ is the rating by the k^{th} annotator for the d^{th} dimension corresponding to the data point n. Armed with this notation, we train a multiple annotator model shown as a graphical model in Figure 1. This model is inspired from the works of Raykar et al. [3] and Gupta et al. [5]. The model assumes that there exists a latent ground truth $\boldsymbol{a}_n^*$ (also of dimensionality D), which is conditioned on the data features. The relationship between the features and $\boldsymbol{a}_n^*$ is captured by the function $f(\boldsymbol{x}_n|\boldsymbol{\theta})$, with parameter $\boldsymbol{\theta}$. We assume f to be an affine projection of the feature vectors as shown in (1), with $\boldsymbol{\theta}$ being the projection matrix.

$$\boldsymbol{a}_n^* = f(\boldsymbol{x}_n|\boldsymbol{\theta}) = \boldsymbol{\theta}^T \begin{bmatrix} \boldsymbol{x}_n \\ 1 \end{bmatrix} \tag{1}$$

The model further assumes that each annotator's ratings are noisy modifications of the ground truth $\boldsymbol{a}_n^*$. We assume these modifications to be the addition of an D-dimensional Gaussian noise with distribution $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, as shown in (2). $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ represent the mean and co-variance matrix of this distribution, respectively.

$$\boldsymbol{a}_n^k = \boldsymbol{a}_n^* + \eta_k, \text{ where } \eta_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{2}$$

2.1. Model training

We estimate the model parameters by maximizing the data log-likelihood. Since the model contains a latent variable (the ground truth $\boldsymbol{a}_*^n$), we adopt the Expectation Maximization algorithm [11] widely used for similar settings. During model training, our objective is to estimate the model parameters $\boldsymbol{\Phi} = \{\boldsymbol{\theta}, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, .., \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K\}$ that maximize the log-likelihood $\mathcal{L}$ of the observed annotator ratings, given the features. Assuming independent data points, $\mathcal{L}$ is given by

$$\mathcal{L} = \log \prod_{n=1}^{N} p(\boldsymbol{A}_n|\boldsymbol{x}_n, \boldsymbol{\Phi}) = \sum_{n=1}^{N} \log p(\boldsymbol{A}_n|\boldsymbol{x}_n, \boldsymbol{\Phi}) \tag{3}$$

The EM algorithm iteratively performs an E-step followed by an M-step. A detailed derivation of these steps for the EM algorithm can be referred from various resources as [11], [4] and [6]. We specifically refer the reader to the EM algorithm derivation in [5] for a multiple annotator model similar to the one presented in this paper. The authors in [5] perform a hard version of EM algorithm where in the E-step an estimate of ground truth $\boldsymbol{a}_n^*$ is computed. This is followed by parameter update in the M-step based on the estimated $\boldsymbol{a}_n^*$. Popular methods such as Viterbi training [13] and K-means clustering [14] are variants of the hard EM algorithm for training Hidden Markov Models and clustering, respectively. Borrowing formulations from the aforementioned research studies, we summarize the E and M steps for obtaining the parameters for the graphical model shown in Figure 1.

EM algorithm

Initialize the model parameters $\boldsymbol{\Phi}$
While the data log-likelihood $\mathcal{L}$ converges, perform
E-step: Estimate the ground truth $\boldsymbol{a}_n^*$ $\forall n = 1..N$ using the optimization stated below. $||.||_2$ represents the l^2-norm in (4).

$$\boldsymbol{a}_n^* = \underset{\boldsymbol{a}_n^*}{\text{argmin}} \sum_{\substack{k=\text{Set of} \\ \text{annotators in } \boldsymbol{A}_n}} \left|\left|\boldsymbol{\Sigma}_k^{-\frac{1}{2}}(\boldsymbol{a}_n^k - \boldsymbol{a}_n^* - \boldsymbol{\mu}_k)\right|\right|_2^2 + \left|\left|\boldsymbol{a}_n^* - \boldsymbol{\theta}^T \begin{bmatrix} \boldsymbol{x}_n \\ 1 \end{bmatrix}\right|\right|_2^2 \tag{4}$$

M-step: Estimate the model parameters $\boldsymbol{\Phi}$ as shown below. N_k is the number of datapoints annotated by annotator k.

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{\substack{n'=\text{Set of datapoints} \\ \text{rated by annotator } k}} \left(\boldsymbol{a}_{n'}^k - \boldsymbol{a}_{n'}^*\right) \tag{5}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k - 1} \sum_{\substack{n'=\text{Set of datapoints} \\ \text{rated by annotator } k}} \Big((\boldsymbol{a}_{n'}^k - \boldsymbol{a}_{n'}^* - \boldsymbol{\mu}_k) * (\boldsymbol{a}_{n'}^k - \boldsymbol{a}_{n'}^* - \boldsymbol{\mu}_k)^T\Big) \tag{6}$$

$$\boldsymbol{\theta} = \underset{\boldsymbol{\theta}}{\text{argmin}} \sum_n \left(\left|\left|a_n^* - \boldsymbol{\theta}^T \begin{bmatrix} \boldsymbol{x}_n \\ 1 \end{bmatrix}\right|\right|_2^2\right) \tag{7}$$

2.2. Model testing

As mentioned before, the goal of our experiments in this work is to predict back the annotator rating, which can be used to address the issue of annotation cost and reducing cognitive load on the annotator by partial prediction of the annotator's ratings. We would like to note that, though our model estimates the latent values for above dimensions, it is hard to evaluate using these since they are unobserved and often subjective in the dataset of interest (as is true for several datasets in the Behavioral Signal Processing domain [15]). In order to predict the rating for the n^{th} file from the k^{th} annotator, we first predict $\boldsymbol{a}_*^n$ following (1) and then add the mean $\boldsymbol{\mu}_k$ of the noise distribution $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, corresponding to the k^{th} annotator. Note that adding $\boldsymbol{\mu}_k$ to $\boldsymbol{a}_n^*$ provides the maximum likelihood estimate of $\boldsymbol{a}_n^k$ thanks to (2) and the Gaussian noise assumption [16].

We use Mean Squared Error (MSE) computed per rating dimension, averaged over all the annotators as our evaluation

Acoustic-prosodic signals	Audio intensity, mel-frequency band, mel-frequency cepstral coefficients and pitch
Statistical functionals	Mean, median, standard deviation, range, skewness and kurtosis

Table 1: Acoustic prosodic signals and their statistical functionals used as features $\boldsymbol{x}_n$ in this study.

metric. For the dimension d (out of D dimensions), we compute the MSE $\mathcal{E}_d$ as shown in (8). I_{nk} is an indicator variable marking if the k^{th} rater annotated the data point n (equation (9)). $a^k_{n,d}$ is the true rating obtained from the rater k on data point n and $\hat{a}^k_{n,d}$ is the model prediction.

$$\mathcal{E}_d = \frac{\sum_{n=1}^{N}\sum_{k=1}^{K} I_{nk}(a^k_{n,d} - \hat{a}^k_{n,d})^2}{\sum_{n=1}^{N}\sum_{k=1}^{K} I_{nk}} \tag{8}$$

where

$$I_{nk} = \begin{cases} 1 & \text{if annotator } k \text{ annotates data point } n \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

We choose this metric as it allows for evaluation on each rating dimension independently. Such a metric is particularly relevant in the Behavioral Signal Processing domain where an evaluation on each dimension of rating is desired. In the next section, we describe the dataset used in this study.

3. Data

We evaluate our model using the SafariBob dataset [12]. The dataset contains multimodal recordings of children watching and imitating video stimuli, each corresponding to a different emotional expression. We extract audio clips from each of these recordings which are annotated over M-Turk. For the purpose of our experiments, we use a set of 244 audio clips (each approximately 25-30 seconds) which were rated over M-Turk by a set of 124 naive annotators. The annotators provide a four dimensional rating ($D = 4$), providing their judgments on expressiveness, naturalness, goodness of pronunciation and engagement of the speaker in each audio clip. The numeric values of these attributes lies in the range of 1 to 5. Each utterance in the data set is annotated by a subset of 15 (out of 124) annotators. This setting is subsumed by the model proposed in section 2. For further details on the dataset, we refer the reader to [12].

3.1. Feature set

We use various statistical functionals computed over a set of acoustic-prosodic properties of the utterance resulting in a set of 474 features ($\boldsymbol{x}_n$) per file. These features are inspired by prior works in speech emotion recognition [17, 18]. The list of the signals and their statistical functionals used as features is shown in Table 1. In the next section, we describe our experimental setup including the baseline model and test different variants of the model described in section 2.

4. Experiments

Based on the approach described in section 2, we train models with different assumptions. Since our goal in these experiments is to predict the annotator ratings, we initially train a baseline system individually modeling every annotator. This is followed by various modifications of the proposed model to predict annotator ratings. We discuss these models in detail below.

4.1. Baseline: Individual annotator modeling

For the baseline, we train individual models for each annotator, instead of the joint model described in section 2. We use an affine projection scheme, for which the relationship between the k^{th} annotator's ratings and features is shown in (10). $\boldsymbol{\theta}_k$ is the projection matrix for the k^{th} annotator. The parameter $\boldsymbol{\theta}_k$ is obtained using minimum mean squared error criterion on the training set, using data points that the annotator rated.

$$\boldsymbol{a}^k_n = f(\boldsymbol{x}_n|\boldsymbol{\theta}_k) = \boldsymbol{\theta}_k^T \begin{bmatrix} \boldsymbol{x}_n \\ 1 \end{bmatrix} \tag{10}$$

4.2. Joint annotator - Independent rating (Joint-Ind) modeling

In this scheme, we train the joint annotator model assuming independence between each dimension in the multidimensional rating. This is achieved by training a separate model for each annotator dimension entry $a^k_{n,d}$. The training procedure is same as presented in section 2, with the special case of ratings being scalar. Consequently, we end up with $D = 4$ different models, one for each dimension. The goal of this model is to identify the benefit of modeling all the annotators jointly, but with an independence assumption between rating dimensions enforced.

4.3. Joint annotator - Joint rating (Joint-Joint) modeling

We next model both the annotators and the ratings jointly as described in section 2. For each annotator we end up with multi-dimensional parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ spanning all four dimensions, which are in turn used to predict the annotator's rating for each data instance. We expect this model to capture any joint relationship between the different dimensions in the ratings, which was not modeled by the previous Joint-Ind model.

4.4. Joint annotator - Conditional rating (Joint-Cond) modeling

The Joint-Cond model is an extension of the model described in section 4.3. In this scheme, we assume partial availability of annotator ratings on a few dimensions. We then use the known distribution parameters for that annotator and the available partial rating to predict the missing dimension. For the sake of brevity we focus on the case when only one of the rating dimensions is missing, noting however that other cases with more than one missing dimension are entirely straightforward. The primary goal of this model is to reduce cognitive load on the annotator by asking him/her to annotate a subset of the rating dimensions.

We represent the available subset of rating dimensions in the vector $\boldsymbol{a}^k_n$, barring rating $a^k_{n,d}$ of dimension d as $\boldsymbol{a}^k_{n,/d}$. Further, we represent the means and co-variance matrix entries corresponding to the dimensions barring dimension d as $\boldsymbol{\mu}^k_{n,/d}$ and $\boldsymbol{\Sigma}^k_{n,/d}$. In our specific case, $\boldsymbol{\mu}^k_{n,/d}$ and $\boldsymbol{\Sigma}^k_{n,/d}$ would be of dimensionalities 3×1 and 3×3, respectively. Also, the entries within $\boldsymbol{\Sigma}^k_n$ storing the co-variances between the dimension d and other dimensions is represented as $\boldsymbol{\Gamma}^k_{n,d}$. $\boldsymbol{\Gamma}^k_{n,d}$ is a vector of dimensionality 1×3. Now, given that the Joint annotator - Joint rating model prediction for the rating at dimension d was given by $\hat{a}^k_{n,d}$, we update it to $\hat{a}^k_{n,d+}$ with the availability of $\boldsymbol{a}^k_{n,/d}$ as shown in (11). This equation follows from the computation of conditional Gaussian distribution from a joint Gaussian distribution, given partial availability of some of the variables [16].

$$\hat{a}^k_{n,d+} = \hat{a}^k_{n,d} + \boldsymbol{\Gamma}^k_{n,d}(\boldsymbol{\Sigma}^k_{n,/d})^{-1}(\boldsymbol{a}^k_{n,/d} - \boldsymbol{\mu}^k_{n,/d}) \tag{11}$$

We report the MSE $\mathcal{E}_d, \forall d \in 1, .., 4$ separately.

5. Results

We report results from two different experiment settings for the models described above. In the first setting, we use ratings from all annotators over the entire data. However, as some of the annotators only annotated a handful of data points (as few as

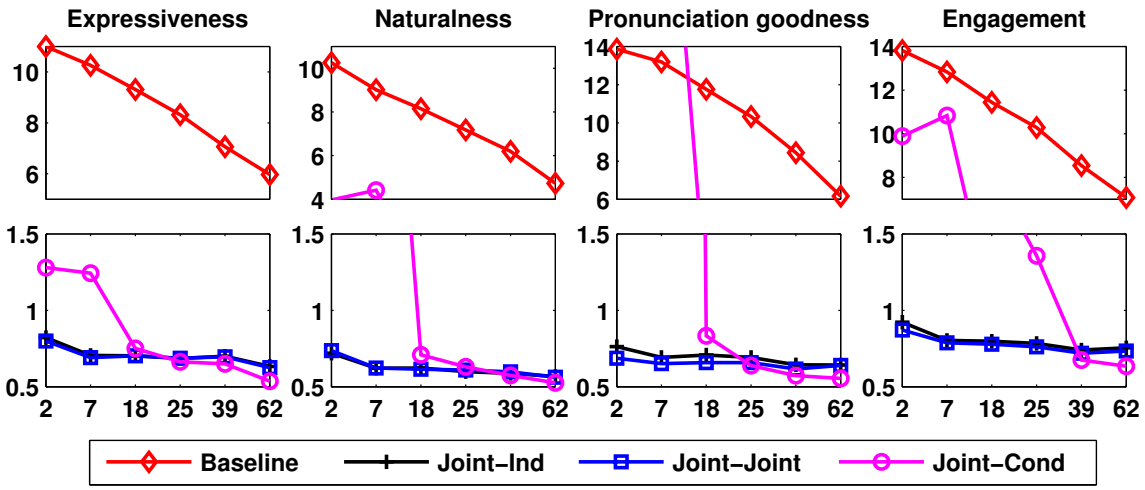

Figure 2: MSE $\mathcal{E}_d$ for the four (baseline, Joint-Ind, Joint-Joint and Joint-Cond) modeling schemes as annotators with less than a threshold count of ratings are dropped. Y-axis represents $\mathcal{E}_d$ and X-axis represents the minimum number of annotations (cutoff threshold).

Dimension d	1 (Ex)	2 (Na)	3 (Go)	4 (En)
Baseline	11.00	10.25	13.86	13.81
Joint-Ind	0.82	0.72	0.76	0.92
Joint-Joint	0.80	0.74	0.69	0.87
Joint-Cond	1.28	3.97	26.08	9.89

Table 2: MSE $\mathcal{E}_d$ for annotator label prediction on the four rating dimensions; Ex: Expressiveness, Na: Naturalness, Go: Goodness of Pronunciation, En: Engagement

2 data points), in the second setting we discard annotators with fewer than a threshold number of ratings. This allows for a more robust estimation of parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ per annotator. We use a 10 fold cross validation scheme over each annotator for all the models.

5.1. Setting 1: Training on data from all annotators

We first compare the different models by including all the annotators in our corpus irrespective of the amount of data they annotated. The metric $\mathcal{E}_d$ for every dimension d is shown in table 2.

From the table, we observe that the Joint-Ind and Joint-Joint models outperform the chosen baseline predictor in all the cases. The Joint-Joint model shows the best performance in 3 out of 4 cases. It makes use of the joint information in the data to make accurate predictions on the annotator ratings rendering confidence in the model's ability to reliably estimate the hidden ground truth along with the model parameters, making this the most desirable model in most cases including when the number of ratings per annotator are low. The Joint-Cond model does better than baseline for expressiveness, naturalness and engagement but fares much worse on pronunciation goodness. We attribute this to poor parameter estimation particularly on annotators with a small number of ratings. In particular the co-variance matrix $\boldsymbol{\Sigma}_k$ is poorly estimated for most annotators, which plays an important role in determining the Joint-Cond estimate. We expect the model to do well when a sufficient amount of rating is available from every annotator, which is discussed in the next section.

5.2. Setting 2: Training on annotators with more than a threshold count of ratings

In this setting, we iteratively remove annotators if they rated fewer than a threshold number of data samples. The metric $\mathcal{E}_d$ is then computed only on the retained annotators. The progression of $\mathcal{E}_d$ as we increase the threshold is shown in Figure 2.

From Figure 2, we observe similar performance trends as the previous section when the cutoff threshold is low. However, as the minimum number of annotations is increased, the baseline and Joint-Cond models show marked improvements in performance, while the Joint-Ind and Joint-Joint models' performance remains more or less consistent. The improvement is significantly better for the Joint-Cond model and it outperforms the Joint-Ind and Joint-Joint beyond a certain threshold for all the rating dimensions. Hence we can use the Joint-Cond model to reduce the dimensionality of queries made to a given annotator, after a sufficient number of ratings are collected for him/her, in turn reducing the annotator's cognitive load and overall annotation cost.

6. Conclusion

Ratings from multiple annotators are often pooled in several applications with a latent ground truth. Several previous works [3] have proposed joint modeling methods for modeling these ratings from multiple annotators. However, such models were not investigated in the case of multidimensional ratings. In this work, we presented a multiple annotator model for multidimensional labels and proposed variants which were applied to the task of predicting back annotator labels. We tested the models on the SafariBob dataset with four dimensional ratings and observed that the proposed models outperformed the baseline and provide mechanisms to make low error label predictions. A further extension was proposed which was shown to be useful in reducing the dimension of ratings presented to annotators after we obtain sufficiently confident parameters.

Future work includes expanding the models to incorporate different types of noises that reflect annotator types such as naive, adversary and/or agnostic. We also plan to expand the model to provide theoretical bounds for prediction errors as a function of the number of data points per annotator. Finally, the model could also be extended to other studies (ex: in the domain of Behavioral Signal Processing) and parameters could be analyzed in light of the domain knowledge for a greater impact.

7. Acknowledgements

The authors gratefully acknowledge support from the National Institutes of Health (NIH-1R01DC012774-01, R21 DC010867-01) and the National Science Foundation (NSF).

8. References

[1] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. Springer-Verlag New York, Inc., 1994, pp. 3–12.

[2] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," in *ACL*, vol. 7, 2007, pp. 440–447.

[3] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *The Journal of Machine Learning Research*, vol. 11, pp. 1297–1322, 2010.

[4] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the em algorithm," *Applied statistics*, pp. 20–28, 1979.

[5] R. Gupta, K. Audhkhasi, Z. Jacokes, A. Rozga, and S. Narayanan, "Modeling multiple time series annotations based on ground truth inference and distortion," [Online; posted]. [Online]. Available: http://sail.usc.edu/~guptarah/smile_paper.pdf

[6] K. Audhkhasi and S. Narayanan, "A globally-variant locally-constant model for fusion of labels from multiple diverse experts without using reference labels," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 4, pp. 769–783, 2013.

[7] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[8] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 160–167.

[9] R. Caruana, S. Baluja, T. Mitchell *et al.*, "Using the future to" sort out" the present: Rankprop and multitask learning for medical risk evaluation," *Advances in neural information processing systems*, pp. 959–965, 1996.

[10] M. L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6965–6969.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[12] R. B. Grossman, L. R. Edelson, and H. Tager-Flusberg, "Emotional facial and vocal expressions during story retelling by children and adolescents with high-functioning autism," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 3, pp. 1035–1044, 2013.

[13] M. Franzini, K.-F. Lee, and A. Waibel, "Connectionist viterbi training: A new hybrid method for continuous speech recognition," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*. IEEE, 1990, pp. 425–428.

[14] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[15] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[16] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[17] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge." in *INTERSPEECH*, vol. 2009. Citeseer, 2009, pp. 312–315.

[18] R. Gupta, C.-C. Lee, and S. S. Narayanan, "Classification of emotional content of sighs in dyadic human interactions," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012.