Bayesian Transfer Learning for the Prediction of Self-reported Well-being Scores

Eirini Christinaki, *Student Member, IEEE*, Riccardo Poli, and Luca Citi, *Senior Member, IEEE*

***Abstract*— Predicting the severity and onset of depressive symptoms is of great importance. User-specific models have better performance than a general model but require significant amounts of training data from each individual, which is often impractical to obtain. Even when this is possible, there is a significant lag between the beginning of the data-collection phase and when the system is completely trained and thus able to start making useful predictions. In this study, we propose a transfer learning Bayesian modelling method based on a Markov Chain Monte Carlo (MCMC) sampler and Bayesian model averaging for dealing with the challenge of building user-specific predictive models able to make predictions of self-reported well-being scores with limited sparse training data. The evaluation of our method using real-world data collected within the NEVERMIND project showed a better predictive performance for the transfer learning model compared to conventional learning with no transfer.**

I. Introduction

Research on constructing models to predict future mood of individuals with depression has shown that, in addition to the expected variables describing the historical mood, the important variables are diverse and individual-dependent [1]. User-specific models provide a better performance than a general model since they are targeted for each specific user. However, user-specific predictions about depressive symptoms require a significant amount of labelled data from the individual subject and the development of appropriate techniques. Traditional machine learning algorithms work under the common assumption that the training and test data are drawn from the same feature space and have the same distribution [2], i.e., that previously collected (labelled or unlabelled) data are similar in nature to future data. Yet, in many real-world applications this assumption does not hold. Therefore, at some point prediction models become unreliable and must be rebuilt and retrained using newly collected training data which is expensive and, sometimes, not practically possible.

This is a problem faced, for instance, in the NEVERMIND [3] project, which aims to provide effective smartphone-based self-management tools to help individuals at risk of developing depressive symptoms as a consequence of a primary disease (e.g., cancer, myocardial infarction, amputation, nephropathy). In this project, sparse multimodal biomedical and subjective data, including a collection of physiological data, body movement, speech, and the recurrence of social interactions, are collected via a smartphone and a lightweight sensorized shirt. The data from individual users are collected over time and become available in a sequential manner. The patient's condition may improve or worsen over time and, so, the feature space and/or the distribution of the data changes from those valid at the time of training.

Another challenge associated with the project is that the entire dataset is not available at once and thus the model needs to be trained incrementally on the data available at a given time. This means that initially the model will be expected to make a prediction for a given individual using a relatively small dataset for that person. When training a model on small datasets with traditional machine learning algorithms, challenges include overfitting, difficulties in handling outliers and differences in the data distribution between training and testing sets.

Transfer learning methods constitute a recent class of techniques overcoming the assumption that the data distributions are the same [4]. These methods can use data from unrelated or partially related tasks [5] and they allow the domains, tasks, and distributions used in training and testing to be different within some limits [6] and without the need of building a new model from scratch. They rely on the basic assumption that the source and target domains are related, implying that an explicit or implicit relationship exists between the feature space of the two domains. The goal of transfer learning is to improve learning in the target domain by leveraging previously acquired knowledge from the source domain. These methodologies have been also employed to improve model accuracy in the presence of scarce data [7]. Transfer learning techniques that take into account population heterogeneity have been proposed in domains involving sequential data modelling [8].

In this paper, transfer learning is used to address the challenge of creating user-specific models and making predictions of self-reported well-being scores when training is performed incrementally on sparse data becoming slowly available over time. We propose a Bayesian transfer learning method that makes predictions about a specific individual by combining models trained on the other participants according to how well they fit his or her past observations.

The rest of this paper is organized as follows: Section II provides a brief description of the non-transfer model used by the NEVERMIND project, introduces the MCMC sampler and describes the proposed transfer learning method. Section III presents the experimental study where the pro-

Eirini Christinaki, Riccardo Poli, and Luca Citi are with the School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom. Email: {e.christinaki, rpoli, lciti}@essex.ac.uk.

This work was supported by the European Commission under the NEVERMIND project (H2020-PHC-689691).

This work is licensed under a Creative Commons Attribution 3.0 License.
For more information, see http://creativecommons.org/licenses/by/3.0/

posed model is evaluated and discussed. Finally, Section IV, provides conclusions and the future research directions.

II. Method

A. Model without transfer

We previously proposed to model subjects and predict their self-reported well-being scores using a Linear Dynamic System (LDS) [9]. The method assumes that the well-being of the user is represented by a state vector, and that its dynamics can be captured by an LDS of the following form:

$$x(t) = Ax(t-1) + Bu(t) + e_x(t) \tag{1a}$$

$$y(t) = Cx(t) + \mu_y + e_y(t) \tag{1b}$$

where $x(t) \in \mathbb{R}^{n_\mathbf{x}}$ is the latent state reflecting the user's well-being condition, $y(t) \in \mathbb{R}^{n_\mathbf{y}}$ is the vector of observations (representing the measurements collected on the user, including biomedical signal features and self-reported well-being scores), $u(t) \in \mathbb{R}^{n_\mathbf{u}}$ is the input vector (representing the influences from the external environment, e.g. weather or day of week), and μ_y is the baseline value of the observation vector. Finally, e_x and e_y represent the state and observation noise, respectively, which are assumed to be distributed as $e_x(t) \sim \mathcal{N}(0, S_x)$ and $e_y(t) \sim \mathcal{N}(0, S_y)$. The LDS model (1) can describe the current state as an auto-regression of arbitrary order simply by extending the state to include its most recent values, e.g. by writing $x(t) = [\xi(t-2), \xi(t-1), \xi(t)]^\top$ where $\xi(t)$ is the original latent state and $x(t)$ is the extended one. The parameters of the LDS model are A, B and C, i.e. the transition, input and observation matrices, respectively. In [9] model fitting, i.e. the identification of such matrices, was accomplished using Expectation Maximization (EM).

B. Detailed Description of Model Inputs

In this work, we use the data collected from 45 participants enrolled in the pilot study of the NEVERMIND project. The observation vector $y(t)$ includes three self-reported well-being scales, namely the numerical answers (provided using a sliding scale from 1 to 6, the lower the better) to the questions "How are you feeling today?", "How was your sleep?" and "How was your day?". Each question is prompted daily and participants may refuse to provide an answer. Subjects that have answered less than 10% of the time on average or those that their total data length was less than two weeks were removed from the analysis carried out in this paper. The inputs $u(t)$ were set to a constant value in this study.

C. Bayesian Transfer Learning

We consider a unit-root third-order autoregressive model, which can be represented by the LDS model (1) with:

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1-a_1-a_2 & a_2 & a_1 \end{bmatrix},\ B = \begin{bmatrix} 0 \\ 0 \\ b \end{bmatrix},\ S_x = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & s_1 \end{bmatrix},$$

$$C = \begin{bmatrix} c_1 & c_2 & c_3 \\ c_4 & c_5 & c_6 \\ c_7 & c_8 & c_9 \\ c_{10} & c_{11} & c_{12} \end{bmatrix},\ S_y = \begin{bmatrix} 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 \end{bmatrix}.$$

The diagonal of the S_y matrix was chosen empirically by estimating the variance of the error made by the subjects when using the slider to provide answers to the questionnaires, also accounting for the fact that the scales were quantized.

While in [9] estimates of the unknown model matrices were obtained using EM, i.e. maximising the likelihood (marginalised over the latent state), here we parametrise them through $\theta = [a_1, a_2, b, c_1, \ldots, c_{12}, s_1]$ and follow a Bayesian approach to obtain their posterior probabilities and perform transfer learning, as explained in the following sections.

1) MCMC sampler: Bayesian inference offers an alternative to maximum likelihood and allows us to determine the posterior probability of the model parameters given the data. From Bayes' theorem, the posterior probability density of a set of model parameters θ given the data D is:

$$p(\theta|D) = \frac{p(D|\theta)\, p(\theta)}{p(D)} \tag{2}$$

where $p(\theta)$ is the prior (our beliefs about the parameters θ before having seen any data D), $p(D|\theta)$ is the likelihood (the probability density of observing the data given a set of parameters), and $p(D)$ is the evidence.

Markov-chain Monte Carlo (MCMC) methods can be used to obtain posterior parameter estimates when these are difficult to express in closed form and works well even for complicated distributions in high-dimensional spaces. MCMC constructs a Markov-chain having as its equilibrium distribution the target posterior distribution. To sample from the posterior distribution of the parameters (our beliefs about the parameters after having seen the data), we used the affine invariant ensemble sampler for MCMC proposed in [10]. Effectively, given a way to compute the product $p(D|\theta)\, p(\theta)$, the ensemble sampler generates random vectors θ distributed according to $p(\theta|D)$.

The likelihood $p(D|\theta)$ is in our case the marginal likelihood of the LDS model in section II-A marginalised over the latent state:

$$p(y|\theta) = \int p(y|x,\theta)\, p(x|\theta)\, \mathrm{d}x, \tag{3}$$

which can be readily obtained using a Kalman filter (see [9]). Additionally, we specify a prior probability distribution $p(\theta)$ to inform and constrain our model. Specifically, we place a Gaussian prior over the c_i coefficients and an inverse gamma prior over the non-zero diagonal element s_1 of S_x. We adopt diffuse priors because they express vague or general information so they are dominated by the likelihood function and have minimal effect, relative to the data, on the final inference.

2) Bayesian Model Averaging: We use transfer learning to tackle the problem of modelling and predicting from limited, sparse sequential data. This information transfer is particularly important in NEVERMIND because by appropriately sharing knowledge between personalised models, there is an opportunity to provide reliable predictions for a given subject even when limited data is available for that subject.

Our approach to transfer learning is based on Bayesian model averaging, whereby the probability of a quantity of

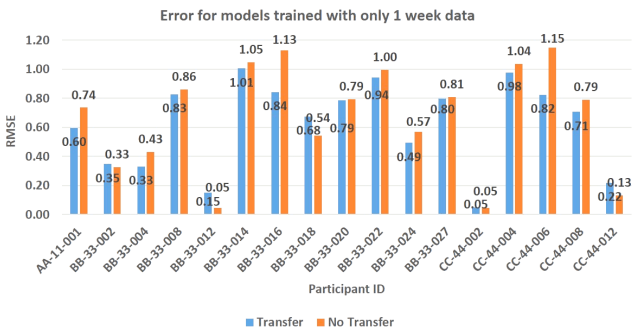

Fig. 1. RMSE per participant for the transfer and non-transfer prediction models trained with 1 week data.

interest Q, such as a future observable, is obtained from multiple models M_k as the average of the posterior distributions under each of the models considered, weighted by their posterior model probability [11]:

$$p(Q|D) = \sum_{k=1}^{K} p(Q|M_k, D)\, p(M_k|D). \tag{4}$$

In our case the quantity of interest is the vector of future self-reported well-being scores y_s^+ for subject s given the past observations y_s collected from the same subject as well as those, $y_{\bar{s}}$, collected from the other subjects. We perform model averaging over a hypothetical continuum of models as follows:

$$\begin{aligned} p(y_s^+|y_s, y_{\bar{s}}) &= \int p(y_s^+|y_s, \theta)\, p(\theta|y_s, y_{\bar{s}})\, \mathrm{d}\theta \\ &= \int p(y_s^+|y_s, \theta)\, \frac{p(y_s|\theta)\, p(\theta|y_{\bar{s}})}{\int p(y_s|\theta')\, p(\theta'|y_{\bar{s}})\, \mathrm{d}\theta'}\, \mathrm{d}\theta \\ &\approx \sum_{k=1}^{K} p(y_s^+|y_s, \theta_k)\, \frac{p(y_s|\theta_k)}{\sum_{j=1}^{K} p(y_s|\theta_j)}, \end{aligned} \tag{5}$$

where the model parameters θ_k in the last expression are distributed according to $p(\theta|y_{\bar{s}})$ and are obtained by drawing random samples from the chains built from the other participants using the MCMC sampler described in II-C.1. The probabilities $p(y_s^+|y_s, \theta_k)$ and $p(y_s|\theta_k)$ can be easily obtained by running a Kalman filter for the LDS model with parameters θ_k (which was obtained using a subject different from s) on the past observations y_s of subject s. In a nutshell, models trained on other subjects are used to make predictions about the future of subject s with each model being weighted by how well it fits the past of subject s.

Calling $\mu_k(t)$ and $\sigma_k^2(t)$ the mean and variance of the future self-reported well being $y_s^+(t)$ as predicted by the k-th model θ_k through the Kalman filter, the mean and variance of the Bayesian model-averaged $y_s^+(t)$ are obtained as follows:

$$\mu(t) = \frac{\sum_k \mu_k(t)\, p(y_s|\theta_k)}{\sum_k p(y_s|\theta_k)}, \tag{6}$$

$$\sigma^2(t) = \frac{\sum_k \left[\sigma_k^2(t) + (\mu_k(t) - \mu(t))^2\right]\, p(y_s|\theta_k)}{\sum_k p(y_s|\theta_k)}. \tag{7}$$

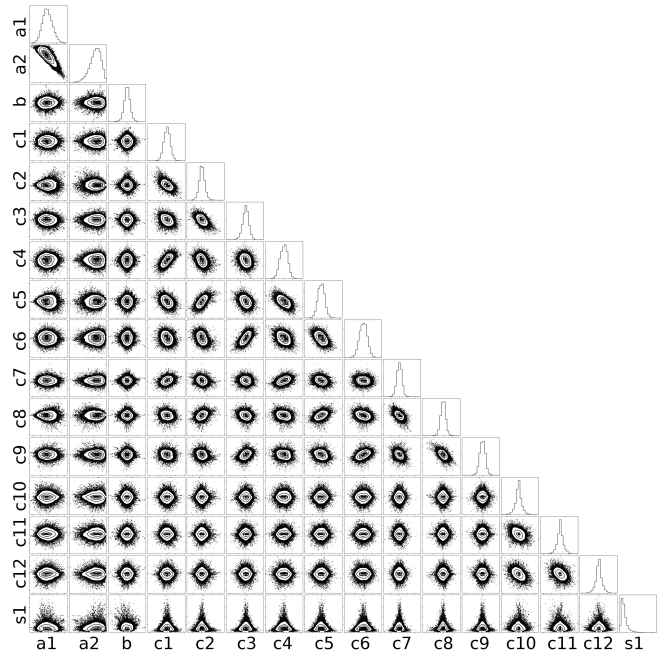

Fig. 2. Corner plot of the 16 parameters of the model. The histograms along the diagonal presents the marginalized distribution for each parameter independently. The other panels show the marginalized two dimensional distributions (the covariance between parameters).

III. Experimental Study

To evaluate our transfer learning model, we assessed its predictive performance and compared it against the non-transfer model that uses the EM method (trained by maximum likelihood). Both models are evaluated using real-world data collected in Pisa, Turin and Lisbon within the NEVERMIND project. The experiments where approved by local ethical committees and all participants have signed an informed consent form.

For the MCMC method, a separate ensemble was trained for each participant s. Each ensemble comprised 130 walkers that sampled our 16-dimensional parameter space for 4,500 iterations, of which the first 1,500 were considered burn-in. The fraction of steps accepted for each walker was around 0.37, which is within the suggested range 20%-50% [10].

The corner plot in Figure 2 shows the one- and two-dimensional projections of the samples obtained by MCMC using the actual self-reported well-being data y_s from one of our participants. These can be interpreted as sampled estimates of the marginal (diagonal plots):

$$p(\theta_i|y_s) = \int p(\theta|y_s)\, \mathrm{d}\theta_1 \ldots \mathrm{d}\theta_{i-1} \mathrm{d}\theta_{i+1} \ldots \mathrm{d}\theta_n$$

and joint (off-diagonal plots) posterior distributions:

$$p(\theta_i, \theta_j|y_s) = \int p(\theta|y_s)\, \mathrm{d}\theta_1 \ldots \mathrm{d}\theta_{i-1} \mathrm{d}\theta_{i+1} \ldots \mathrm{d}\theta_{j-1} \mathrm{d}\theta_{j+1} \ldots \mathrm{d}\theta_n.$$

In this section, we present outcomes from 17 randomly selected test subjects. Initially, for each subject s, the models were trained with 7 days (1 week) of historic data y_s from the same subject to predict the observations y_s^+ of the next

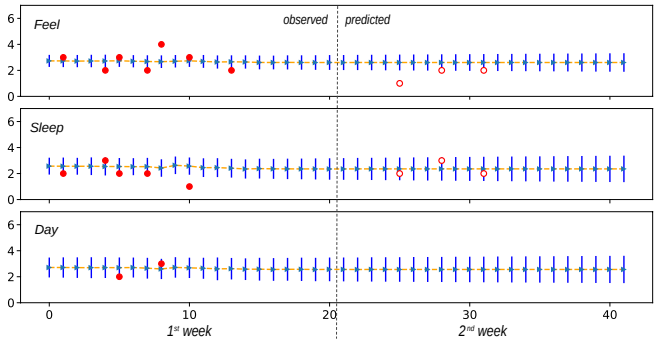

Fig. 3. Example of self-reported well-being score modelling and prediction. The model was trained with one week of data (the leftmost 21 points at 8-hour resolution) and tested on the following week. The solid red circles represent the reported scores that were used by the model, while the empty ones in the second week are only reported for reference. The blue triangles and the associated whiskers represent the mean and standard deviation, respectively, of the state predicted by the model according to (6) and (7).

TABLE I

COMPARISON OF PREDICTION RESULTS FOR THE TRANSFER AND NO-TRANSFER LEARNING MODELS

	RMSE		*Accuracy [%]*	
Data Length	*Transfer*	*No transfer*	*Transfer*	*No transfer*
1 week	**0.62**	0.67	**62.74**	60.55
2 weeks	**0.61**	0.74	**63.93**	56.62
3 weeks	**0.66**	0.77	**52.44**	49.81
4 weeks	**0.55**	0.91	**61.45**	34.05
Avg.	**0.61**	0.77	**60.14**	50.26

Bold values show the best evaluation scores (lowest error, highest accuracy) among the predictive models.

7 days (test week). The predictions were then compared with the actual observations for the test week. For comparing and quantifying the prediction ability of each model we used metrics including accuracy and root mean square error (RMSE). The accuracy was computed considering a prediction as correct if the predicted value was less than 0.5 away from the actual observation.

The results presented in Figure 1 show that the transfer learning model yields the lowest RMSE in all cases. The overall RMSE for this scenario was 0.62 for our method and 0.67 for the no-transfer one. The accuracy measured from the average predicted results was 62.74% for the transfer learning approach and 60.55% for the no-transfer EM method. An indicative example of the predicted mean and the variance as learned by our model along with the sparse self-reported well-being scores from one of our participants can be seen in Figure 3.

To investigate how the predictive performance changes as more data become available, we trained the models for each participant with 14 days (2 weeks), 21 days (3 weeks) and 28 days (4 weeks) of historic data to predict the following 7 days (test week). The average RMSE and accuracy for the experiments performed with varying training data length are presented in Table I. The evaluation metrics for the predictive models show that the best scores were obtained using the transfer learning method in all four scenarios.

IV. CONCLUSION

In this paper we propose a transfer learning Bayesian modelling method based on an MCMC sampler and Bayesian model averaging to deal with the challenge of building user-specific predictive models able to make predictions with limited sparse training data. According to our experimental results, our model performs better than training separate models for each participant by using solely their examples. Its overall performance shows the advantage of delivering better results for participants with very few training samples. Our method adequately deals with the uneven sparse data representation in the dataset and produces a better suited model for participants with very few training samples.

In future research, we plan to work towards clustering our data and defining patients' groups with similar characteristics. This could be achieved by defining clusters of participants with e.g. same gender, age group, and/or primary diseases. These clusters will be used to place a prior on the random selection of model parameters so that models from subjects with similar characteristics will have a higher probability of being selected.

ACKNOWLEDGEMENT

The authors would like to thank Dr Xinyang Li for helping with data pre-processing, the NEVERMIND [3] consortium partners for providing the data and the relevant infrastructure and finally, the participants who took part in this study.

REFERENCES

[1] J. Pastor and W. Van Breda, "Analyzing and Predicting Mood of Depression Patients," 2015. [Online]. Available: https://science.vu.nl/en/Images/werkstuk-pastor{_}tcm296-664630.pdf

[2] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, oct 2010.

[3] "NEurobehavioural predictiVE and peRsonalised Modelling of depressIve symptoms duriNg primary somatic Diseases with ICT-enabled self-management procedures," Online at http://www.nevermindproject.eu/, 2018.

[4] R. Sousa, L. M. Silva, L. A. Alexandre, and J. Santos, "Transfer Learning: Current Status, Trends and Challenges," in *20th Portuguese Conference on Pattern Recognition, RecPad*, 2014, pp. 57–58.

[5] D. M. Roy and L. P. Kaelbling, "Efficient Bayesian Task-Level Transfer Learning," *IJCAI'07 Proceedings of the 20th international joint conference on Artifical intelligence*, pp. 2599–2604, 2007.

[6] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang, "Transfer learning using computational intelligence: A survey," *Knowledge-Based Systems*, vol. 80, pp. 14–23, 2015.

[7] A. Maxhuni, P. Hernandez-Leal, L. E. Sucar, V. Osmani, E. F. Morales, and O. Mayora, "Stress modelling and prediction in presence of scarce data," *Journal of Biomedical Informatics*, vol. 63, pp. 344–356, 2016.

[8] P. Jaini, Z. Chen, P. Carbajal, E. Law, L. Middleton, K. Regan, M. Schaekermann, G. Trimponias, J. Tung, and P. Poupart, "Online Bayesian Transfer Learning for Sequential Data Modeling," in *International Conference on Learning Representations (ICLR)*, 2017.

[9] X. Li, R. Poli, G. Valenza, E. P. Scilingo, and L. Citi, "Self-reported well-being score modelling and prediction: Proof-of-concept of an approach based on linear dynamic systems," in *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2017, pp. 2205–2208.

[10] J. Goodman and J. Weare, "Ensemble Samplers With Affine Invariance," *Communications in Applied Mathematics and Computational Science*, vol. 5, no. 1, pp. 65–80, 2010.

[11] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian Model Averaging: A Tutorial," *Statistical Science*, vol. 14, no. 4, pp. 382–401, 1999.