Multi-task Learning for Predicting Health, Stress, and Happiness

Natasha Jaques,* **Sara Taylor***, **Ehimwenma Nosakhare, Akane Sano, Rosalind Picard**
MIT Media Lab Affective Computing Group
Cambridge, MA 02139
{jaquesn, sataylor, ehinosa, akanes, picard}@media.mit.edu

Abstract

Multi-task Learning (MTL) is applied to the problem of predicting next-day health, stress, and happiness using data from wearable sensors and smartphone logs. Three formulations of MTL are compared: i) Multi-task Multi-Kernel learning, which feeds information across tasks through kernel weights on feature types, ii) a Hierarchical Bayes model in which tasks share a common Dirichlet prior, and iii) Deep Neural Networks, which share several hidden layers but have final layers unique to each task. We show that by using MTL to leverage data from across the population while still customizing a model for each person, we can account for individual differences, and obtain state-of-the-art performance on this dataset.

1 Introduction

Perceived wellbeing, as measured by self-reported health, stress, and happiness, has a number of important clinical health consequences. Self-reported health is not only strongly related to actual health, but to all-cause mortality [12]. Stress increases susceptibility to infection and illness [4]. Finally, not only is self-reported happiness indicative of scores on clinical depression measures [3], but happiness is so strongly associated with greater longevity that the effect size is comparable to that of cigarette smoking [20]. The ability to model and predict these measures could therefore be immensely beneficial in the treatment and prevention of both mental illness and disease.

Unfortunately, modeling wellbeing and mood has historically been a difficult task, with typical classification accuracies ranging from 55-80% (e.g. [1,5,11,13]), even with sophisticated models or multi-modal data. In this paper, we use a challenging dataset where accuracies from prior efforts to recognize wellbeing and mood ranged from 56-81% [7,8]. Further, almost all work previous work relates to detection rather than prediction, and is based on data gathered in a lab.

Multi-task Learning (MTL) is a type of transfer learning, in which models are learned simultaneously for several related tasks, but share information through similarity constraints [2]. By using MTL to account for individual differences in the relationship between behavior and wellbeing, we are able to obtain groundbreaking performance on this dataset (82-87% accuracy) , significantly improving on prior published results. Our data are gathered in the "wild" as participants go about their daily lives, and we are able to predict participants' future wellbeing, as opposed to simply detecting their current state. Further, we provide novel clinical insights through an implicit clustering of users and weighting of input sources. These innovative ideas can greatly improve future modeling of health, stress, and happiness.

2 Dataset and Classification Problem

The data for this research were collected as part of a long-term study of undergraduate students entitled SNAPSHOT[1] [17]. Participants were monitored for 30 days each, during which wearable

*Both authors contributed equally to this work
[1]Sleep, Networks, Affect, Performance, Stress, and Health using Objective Techniques

sensor, smartphone, and survey data were collected. Students self-reported their health, stress, and happiness in the morning and evening every day on a scale from 0-100. Binary classification labels were assigned to the top and bottom 30% of these scores [7, 8].

A total of 343 Features were extracted from sensors, smartphones logs, behavioral surveys, and weather information. Details are provided in [7,8,18]; here we provide a brief overview. Wrist-worn Affectiva Q sensors were used to collect 24-hour-a-day skin conductance (SC), skin temperature, and 3-axis accelerometer data, from which features such as step count, stillness, and skin conductance responses (which relate to emotional arousal and stress) were extracted. A smartphone app logged participants' calls, text messages, screen on/off events, and location throughout the day. In addition to extracting features about participants' communication and phone usage, location patterns were modeled with a Gaussian Mixture Model.

Daily survey features included self-reported behaviors such as academic activities, exercise, and sleep. Weather features were computed using data from DarkSky's Forecast.io API [14]. Finally, we included 3 extrinsic variables available to any smartphone app: the participant ID, the day of the week, and whether it is a night before a school day (when they are likely to have more workload).

MTL was applied to this data in two ways: 1) we treat predicting each wellbeing label (*health, stress*, and *happiness*) as one task, and 2) we treat predicting the wellbeing of a single user as one task. For the second *users-as-tasks* formulation, we discarded users with less than 10 days worth of data, since they could not provide enough data to train viable models, and performed feature selection to limit overfitting. For the wellbeing-as-tasks case we obtained a dataset of 187 users, 343 features and 1071 labeled data points; for users-as-tasks there were 83 users, 27 features, and 878 labeled data points. A random 80/20% split was used to partition the data into training and test sets. 5-fold cross validation was applied to the training set to find optimal parameter settings.

3 Methods

3.1 Multi-Task Multi-Kernel Learning (MTMKL)

Multi-Task Multi-Kernel Learning (MTMKL) is a modified version of MKL in which tasks share information through kernel weights [10]. An RBF kernel was applied to features from each of eight data modalities: physiology, location, call, SMS, screen, survey, weather, and extrinsic. These values are then combined into a single kernel function for each task, $k_{\boldsymbol{\eta}}^{(t)}$, via a weighted sum parameterized by the task's modality weights, $\boldsymbol{\eta}^{(t)}$. The optimal $\boldsymbol{\eta}^{(t)}$ for each task can be learned by maximizing an objective function similar to that of a Support Vector Machine (SVM) with a least squares loss function (LSSVM) and kernel $k_{\boldsymbol{\eta}}^{(t)}$, but with constraints $\boldsymbol{\eta}_m^{(t)} \geq 0 \ \forall m$ and $\sum_{m=1}^{M} \boldsymbol{\eta}_m^{(t)} = 1$. In order for information to be shared among tasks, task weights are regularized globally by penalizing divergence from other weights. An iterative gradient descent method proposed in Kandemir et al. [10] was used to train the model.

Note that an LSSVM must be trained for each task. This is impossible if there is only one classification label in the training data; e.g. if a user was consistently healthy during the randomly selected training days. Thus for the users-as-tasks formulation we apply K-means to participants' pre-study survey data[2] in order to cluster the participants, and treat predicting the wellbeing of each cluster as one task. The optimal number of clusters was found to be 2 based on silhouette score [16].

3.2 Hierarchical Bayes with Dirichlet Process Priors (HBDPP)

Hierarchical Bayesian learning is another popular approach to MTL, in which a shared prior is placed on the model parameters for each task, constraining the tasks to be similar (e.g. [15] [21]). Since it is assumed that tasks are related to one another, it is ideal to both allow similar tasks to share information, and to identify tasks that are very similar to one another. Here, we employ a non-parametric Bayesian hierarchical model [22] to (i) cluster related users and (ii) perform MTL by jointly learning logistic regression classifiers for each cluster.

Specifically, the model draws logistic regression weights w_m for each task from a common prior distribution, G, sampled from a Dirichlet process with a scaling parameter $\alpha > 0$ (an innovation

[2] Including GPA, personality, anxiety, mental and physical health, sleep quality, and perceived stress.

parameter that controls the probability of creating new clusters), and a base d-dimensional multivariate normal distribution G_o. Data from all tasks contribute to the learning of the common prior, and information is transferred via sufficient statistics. From the Chinese restaurant representation of the Dirichlet process, we see that the DP prior induces a clustering effect over the w_m parameters and therefore implicitly clusters the M tasks. Thus, there are only K distinct sets of weights, w_k, one for each cluster. For each task, the model learns the probability that the task belongs to cluster k, π_k. A one-hot encoding of each task's cluster assignment, $c_{m,k}$, selects the appropriate cluster weights w_k for a given task. Mean-field variational Bayesian inference is used to learn the posterior distribution of the latent variables given the observed data and the hyperparameters [6].

3.3 Neural Networks (NNs)

Our MTL network architecture shares some number of hidden layers between all tasks, while the final layers are task-specific (Figure 1 shows a simplified diagram of this design). The net is trained iteratively, by taking a batch of data from a randomly selected task and updating both the shared weights and the task-specific weights.

Because of the small dataset size, we employ several regularization techniques beyond multi-tasking, including imposing a penalty on the L2 norm of the network weights, and applying dropout [19]. Based on previous work that has successfully trained MTL NNs with few samples [9], we choose a simple, fully-connected design with 2-3 hidden layers.

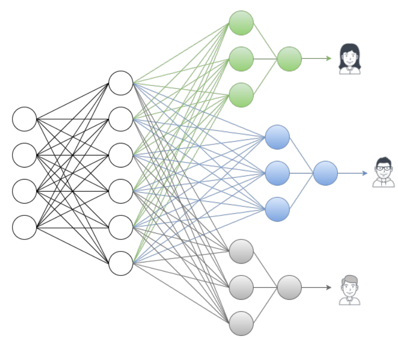


Figure 1: A simplified version of the MTL neural network architecture.

3.4 Single task learning (STL) techniques

To understand the impact of MTL over STL, we implement the STL equivalents of each of the above algorithms: LSSVM (MTMKL), Logistic Regression (LR) (HBDPP), and single-task NNs.

4 Results

The performance of both the STL and MTL algorithms are shown in Table 1. For wellbeing-as-tasks, multi-tasking benefits NNs for both the happiness and stress tasks, but not for health. Since happiness and stress are more strongly correlated with each other than either is with health, the NN likely learned representations that were more related to these than health. The performance of HBDPP and MTMKL generally suffered for the wellbeing-as-tasks problem. One explanation is that the multi-tasking constraints are causing these models to underfit the data. If the model is not sufficiently complex, the regularization effect of MTL may reduce the power of the models to fit the optimal decision boundary. The training accuracy for these models was low, lending support to this theory. In contrast, the NN is a sufficiently complex model that the regularizing force of the additional tasks provides a benefit in terms of generalization ability.

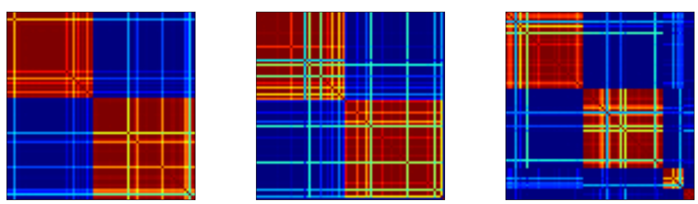
	Classifier	Happiness	Stress	Health
STL	LSSVM	62.27%, .6266	66.98%, .6703	66.05%, .6525
	LR	64.38%, .6832	65.30%, .6839	62.56%, .6609
	NN	62.04%, .6193	64.65%, .6453	67.67%, .6761
MTL - wellbeing	MTMKL	60.19%, .6161	63.97%, .6456	63.26%, .6366
	HBDPP	59.44%, .5943	60.20%, .6035	62.02%, .6236
	NN	68.95%, .6886	69.41%, .6937	64.84%, .6469
MTL - users	MTMKL	75.87%, .7608	75.12%, .7470	75.48%, .7493
	HBDPP	82.17%, .8220	86.07%, .8595	85.58%, .8507
	NN	82.52%, .8266	85.57%, .8541	87.50%, .8682

Table 1: Prediction performance (Accuracy and AUC) of the STL, MTL-wellbeing, and MTL-user methods. Multi-tasking over people is better for all three MTL formulations.

As is evident in Table 1, multi-tasking over users provides markedly increased performance, with accuracies improving by 11-16% across different model types. The improved performance is not due to the reduced feature set size, or the reduced dataset; we tested the wellbeing-as-tasks algorithms with both types of data and found that they performed similarly or worse; in some cases the accuracy dropped by as much as 8%. This is reasonable, given the reduced dataset size. It is also worth noting that both the STL and MTL-wellbeing models had access to the participant ID as a feature, so it is not simply providing this information that boosts the performance of the MTL-user models. The performance improvement resulting from the users-as-tasks formulation therefore suggests that accounting for individual differences may be critical to accurately predicting wellbeing. While it would not be possible to train a reasonable model such as a NN to predict a person's wellbeing from only a few days of data, MTL allows each person to have a model tuned specifically for them, while still learning from data from the rest of the participants.

As described in Section 3.1, MTMKL learns optimal kernel weights for each modality. Because a study of this magnitude is taxing on participants and expensive to conduct, these kernel weights can give pivotal insight into what data are important to collect for predicting wellbeing. For the wellbeing-as-tasks problem, physiology and survey features received the most weight, suggesting SC, movement, lifestyle (e.g. caffeine intake), and sleep are highly relevant to overall wellbeing.

An implicit clustering of tasks is learned by the HBDPP algorithm; Figure 2 plots the probability with which two users will be placed in the same cluster for the health, stress, and happiness tasks. These clusters provide insight into groups of people who have a different relationship between the behavioral features and their resulting wellbeing. For example, a cluster in the happiness model which had only 42% happy days, also had lower GPA, agreeableness, mental and physical health, and higher neuroticism. Another had the highest proportion of happy days, at 69%, and interestingly had low GPA and conscientiousness, but high agreeableness. This cluster also spent less time indoors, slept longer, texted more, had more positive social interactions, and less routine location patterns. HBDPP identified these clusters without access to any information about participant's pre-study personality test results. The clusters for stress and health also revealed differences in personality, sleeping habits, and mental health. Taken together, these findings shed light on how personality, lifestyle, and mental health mediate the effects of these features on health and wellbeing.



(a) Health (2 clusters) (b) Stress (3 clusters) (c) Happiness (4 clusters)

Figure 2: HBDPP participant clusters for the 3 users-as-tasks classification problems. The i, jth cell represents the probability that person i is in the same cluster as person j. The more red (blue) the cell, the more (less) likely that pair of people will be grouped into the same cluster.

5 Conclusions and Future Work

This work has demonstrated that accounting for individual differences through MTL can dramatically improve wellbeing prediction performance. The improvement is not simply due to the application of MTL, but rather to the ability of MTL to allow each user to have a model trained specifically for them, which also benefits from the data of other users. The three formulations of MTL we have explored offer different strengths, including the ability to learn how the importance of feature types differs across tasks and the ability to learn implicit groupings of users.

A major contribution of this research is enabling data-driven insights about behaviors that may contribute to poor wellbeing and declines in mental health, or to a happy, calm, and healthy state. By examining the models' weights and clusters, we can identify behaviors that may be significant predictors for each person, while accommodating different personalities and lifestyles. Hypotheses related to these behaviors can subsequently be tested via causal inference techniques. Traditional methods that derive results from models used to fit groups may or may not apply to individuals. In contrast, the methods shown here have the potential to leverage group data, while also identifying behaviors that may truly influence each individual's path to better health and wellbeing.

References

[1] A. Bogomolov et al. Daily stress recognition from mobile phone data, weather conditions and individual traits. In *Int. Conf. on Multimedia*, pages 477–486. ACM, 2014.

[2] R. Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[3] H. Cheng and A. Furnham. Personality, self-esteem, and demographic predictions of happiness and depression. *Personality and individual differences*, 34(6):921–942, 2003.

[4] S. Cohen et al. Psychological stress and susceptibility to the common cold. *New England journal of medicine*, 325(9):606–612, 1991.

[5] S. D'Mello and A. Graesser. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *UMUAI*, 20(2):147–187, 2010.

[6] Z. Ghahramani and M.J. Beal. Propagation algorithms for variational bayesian learning. *Advances in neural information processing systems*, pages 507–513, 2001.

[7] N. Jaques et al. Multi-task, multi-kernel learning for estimating individual wellbeing. In *NIPS 2015 Workshop on Multimodal ML*, volume 898, 2015.

[8] N. Jaques et al. Predicting students' happiness from physiology, phone, mobility, and behavioral data. In *ACII*. IEEE, 2015.

[9] F. Jin and S. Sun. Neural network multitask learning for traffic flow forecasting. In *IEEE Int'l Joint Conf. on Neural Networks*, pages 1897–1901. IEEE, 2008.

[10] M. Kandemir et al. Multi-task and multi-view learning of user state. *Neurocomputing*, 139:97–106, 2014.

[11] A. Kapoor and R. Picard. Multimodal affect recognition in learning environments. In *ACM Int'l Conf. on Multimedia*, pages 677–682. ACM, 2005.

[12] A. Keller et al. Does the perception that stress affects health matter? the association with health and mortality. *Health Psychology*, 31(5):677, 2012.

[13] R. LiKamWa et al. Moodscope: building a mood sensor from smartphone usage patterns. In *Int. Conf. on Mobile systems, applications, and services*, pages 389–402. ACM, 2013.

[14] The Dark Sky Company LLC. Dark sky forecast api, 2016.

[15] M. Rosenstein et al. To transfer or not to transfer. In *NIPS 2005 Workshop on Transfer Learning*, volume 898, 2005.

[16] P. J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[17] A. Sano. *Measuring College Students Sleep, Stress and Mental Health with Wearable Sensors and Mobile Phones*. PhD thesis, MIT, 2015.

[18] A. Sano et al. Prediction of happy-sad mood from daily behaviors and previous sleep history. In *EMBC*. IEEE, 2015.

[19] N. Srivastava et al. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[20] Ruut Veenhoven. Healthy happiness: Effects of happiness on physical health and the consequences for preventive health care. *Journal of happiness studies*, 9(3):449–469, 2008.

[21] A. Wilson et al. Multi-task reinforcement learning: a hierarchical bayesian approach. In *ICML*, pages 1015–1022. ACM, 2007.

[22] Y. Xue et al. Multi-task learning for classification with dirichlet process priors. *The Journal of Machine Learning Research*, 8:35–63, 2007.