Multi-task, Multi-Kernel Learning for Estimating Individual Wellbeing

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

Abstract

We apply a recently proposed technique – Multi-task Multi-Kernel Learning (MTMKL) – to the problem of modeling students' wellbeing. Because wellbeing is a complex internal state consisting of several related dimensions, Multi-task learning can be used to classify them simultaneously. Multiple Kernel Learning is used to efficiently combine data from multiple modalities. MTMKL combines these approaches using an optimization function similar to a support vector machine (SVM). We show that MTMKL successfully classifies five dimensions of wellbeing, and provides performance benefits above both SVM and MKL.

1 Introduction

Depression is a widespread and serious problem, disproportionately affecting college-aged individuals [20]. The ability to handle negative life events without becoming depressed, termed *resilience*, depends on several factors related to overall wellbeing; these include social support, engagement with work, happiness [18], physical health [4], and sleep [19].

To better understand the factors that contribute to wellbeing and resilience among college students, researchers at MIT and Brigham and Women's Hospital are conducting an ongoing study called SNAPSHOT: Sleep, Networks, Affect, Performance, Stress, and Health using Objective Techniques [14]. As the name suggests, the study gathers a massive amount of data about the participants, including five measures of their wellbeing: stress, health, energy, alertness, and happiness.

Previous attempts to analyze this data have looked at individually classifying academic performance, stress, sleep, mental health and happiness [7], [15], [16]. While studying these individual components is informative, the true goal of the research is to be able to model and predict a student's overall wellbeing: a hidden, internal state consisting of a number of the factors mentioned above. The purpose is both to gain insight into factors that could potentially improve students' wellbeing, but also to detect when a student's wellbeing is suffering and therefore help direct intervention efforts.

This work applies Multi-Task Multi-Kernel Learning (MTMKL) [8], to the problem of modeling the five components of wellbeing collected from the SNAPSHOT study. We show that MTMKL is well suited to this problem because of its ability to intelligently combine data from multiple modalities, and share this information among multiple related tasks.

2 Related Work

Numerous studies have used multi-modal data to classify affective states that relate to wellbeing, such as happiness and stress (e.g., [1], [17]). However, it remains a difficult problem, with classifi-

---

∗Both authors contributed equally to this work

cation accuracies typically ranging from 65-72% [1] [7] [10] when modeling ambiguous outcomes like mood, happiness, or stress.

Multi-view learning methods study how to effectively combine data from multiple disparate sources [21], making them ideal for Affective Computing studies with data from multiple modalities. A *view* in this case refers to a subset of related features; for example, one view could be defined as the features from a physiological sensor, while another contains features collected with a smartphone. A common approach to multi-view learning is Multiple Kernel Learning (MKL) [6], in which each view or modality is represented by its own kernel function.

Because our application involves classifying multiple highly related outputs (i.e. the five related measures of wellbeing), we can also benefit from Multi-task learning (MTL). Multi-task learning is a special type of transfer learning [11], in which a classifier is trained on several related tasks at once, while sharing information across the tasks [3]. It is particularly useful if each task is a one-dimensional aspect of a more complex unknown state of the user [8]; for example, related dimensions of wellbeing. MTL can also be applied by defining a 'task' as classifying the data from a particular user, and thus allowing the model to better accommodate sindividual differences.

Recent work by Kandemir [8] has extended and combined these two techniques in a method dubbed Multi-Task Multi-Kernel Learning (MTMKL). Broadly, the classifier combines the kernels for each modality using a set of kernel weights for each task. These weights are regularized globally, allowing information about the weighting to be shared across tasks.

3 Dataset

This dataset was collected from 68 students over 30 days each, resulting in over 2000 days of physiological, survey, and smartphone data. The extensive data makes feature extraction and selection critically important. Since this process has been described in detail in previous work [7], here we simply give a brief overview. Each feature was computed at the granularity of one day, defined from 12:00am to 11:59pm. Features were selected in order to optimize validation accuracy in classifying self-reported happiness, using an iterative process involving both Wrapper Feature Selection (WFS) [5], which tests which subsets of features work most effectively with the desired classifier, and filtering based on Information Gain. The eventual dataset contains 6, 81, 15, and 98 features in the data modalities mobility, smartphone, survey, and physiology, respectively (this was reduced from original feature set sizes of 15, 289, 32, 426).

3.1 Mobility

During the study, an app on the participants' smartphones logged their GPS coordinates throughout the day, which were used to compute a variety of features; for example, distance traveled, the radius of the minimal circle enclosing the subject's location samples, and time spent on the university campus [7]. Whether the phone's data signal came from WiFi or GPS was used to approximate the time the participant spent outdoors. Finally, a Gaussian Mixture Model (GMM) was used to induce a probability distribution over each participant's locations, and thereby compute features related to the likelihood of the location pattern of each day.

3.2 Smartphone

The smartphone logs provided information about the participant's communication over phone and SMS, as well as each time the phone screen was turned on or off. Features related to the timing, duration, and frequency of these activities were computed, as well as the number of unique contacts with whom the participant interacted.

3.3 Survey

Participants submitted daily self-reports on their academic activities, extra-curricular activities, exercise, sleeping, napping, studying, social interactions, and alcohol and drug consumption.

3.4 Physiology

Participants wore an Affectiva Q sensor, which collected skin conductance (SC), skin temperature, and acceleration at 8Hz. Because skin conductance relates to activity of the Sympathetic Nervous System (SNS) – the body's "fight or flight" response – it is frequently used in studies related to emotion and stress [2]. Similarly, acceleration gives an indication of the participant's physical activity, which is strongly related to wellbeing [12]. A variety of signal processing and algorithmic techniques were applied to the raw SC and accelerometer data to extract over 400 features, which were then pruned using feature selection; see [7] for details.

3.5 Classification labels

Students in the SNAPSHOT study self-report their happiness, health, alertness, energy, and stress using a scale from 0-100 every day in the morning and evening. In previous work [7], a binary classification task for happiness was created by labeling the 30% of days which had the highest evening happiness score as positive samples, and the 30% of days with the lowest evening score as negative samples, then randomly partitioning this data into a train, validation and test set. We use the same datasets to facilitate comparison with previous work, but extend them by creating additional labels for the other 4 wellbeing measures. For a given day in the dataset, if a measure such as 'health' had a score higher than the median[1] for this measure, it was labeled as a positive example; otherwise, it was labeled as a negative example. Using this method we are able to retain all of the original data points. Classifying each of the 5 measures is treated as one task in the MTMKL classifier. The final dataset is composed of 5 tasks, 200 features, and 768 data points[2].

4 Methods

4.1 Multiple kernel learning

As in typical MKL, we assign a kernel to each set of features that make up a modality. Here we restrict the model space to using the same kernel function for each modality. In particular, we represent each modality m by one kernel k_m. These are combined into a single kernel, $k_{\boldsymbol{\eta}}$, in a convex combination parameterized by the modality weighting, $\boldsymbol{\eta}$. Noting that $\mathbf{x}_i^{(m)}$ is the ith feature vector that contains only the features belonging to modality m, and M is the total number of modalities, $k_{\boldsymbol{\eta}}$ is defined as follows:

$$k_{\boldsymbol{\eta}}(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\eta}) = \sum_{m=1}^{M} \boldsymbol{\eta}_m k_m(\mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)})$$

The optimal $\boldsymbol{\eta}$ can be learned from the data by maximizing an objective function similar to that of the traditional support vector machine (SVM), but with the multiple kernel, $k_{\boldsymbol{\eta}}$, as defined above, with the constraints that $\boldsymbol{\eta}_m > 0 \; \forall m$ and $\sum_{m=1}^{M} \boldsymbol{\eta}_m = 1$. Noting that $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise, $\boldsymbol{\alpha}_i$ is the ith dual coefficient term, and N is the total number of data points, we write the objective function for the MKL method as follows:

$$\max_{\boldsymbol{\alpha}} J(\boldsymbol{\alpha}, \boldsymbol{\eta}) = \max_{\boldsymbol{\alpha}} \left[\sum_{i=1}^{N} \boldsymbol{\alpha}_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_j y_i y_j \left(k_{\boldsymbol{\eta}}(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\eta}) + \frac{\delta_{i,j}}{2C} \right) \right]$$

4.2 Multi-task Multi-Kernel Learning (MTMKL) machines

Following Kandemir et al. [8], we formulate the multi-task multi-kernel learning (MTMKL) method over T tasks, such that each task t learns a different weighting, $\boldsymbol{\eta}^{(t)}$, on the modalities. These weights are regularized globally by penalizing divergence. Specifically, the objective function Kandemir has proposed minimizes the maximum of the sum of the regularizer term $\Omega(\{\boldsymbol{\eta}^{(t)}\}_{t=1}^{T})$ and the sum of the objective functions of all task learners, $J_t(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\eta}^{(t)})$:

---

[1]The median for each measure is computed over all participants from the evening self-reports.

[2]The size of the dataset is reduced because of missing data

$$\min_{\boldsymbol{\eta}} \left[\max_{\boldsymbol{\alpha}} \left(\Omega(\{\boldsymbol{\eta}^{(t)}\}_{t=1}^T) + \sum_{t=1}^T J_t(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\eta}^{(t)}) \right) \right]$$

For simplicity of notation, we let the inner optimization problem be $O_{\boldsymbol{\eta}}$ such that the objective function of MTMKL becomes $\min_{\boldsymbol{\eta}} O_{\boldsymbol{\eta}}$.

In this paper we focus on two regularization functions, namely $L1$: the inner product regularizer, which penalizes negative total correlation between task weights and $L2$: the l_2-norm regularizer, which penalizes the Euclidean distance between task weights [8]. In both regularizers, ν is used to weight the importance of the divergence. In particular, when $\nu = 0$ the tasks are treated independently and as ν increases, the tasks weights become increasingly restricted to being the same.

We then use the iterative gradient descent method proposed by Kandemir to train the model given the model parameters [8]. This method iterates between (1) solving an SVM for each task given $\boldsymbol{\eta}^{(t)}$ and (2) updating $\boldsymbol{\eta}$ in the direction of $-\partial O_{\boldsymbol{\eta}}/\partial \boldsymbol{\eta}^{(t)}$ (see Algorithm 1). Using the constraints on the kernel function and noting $\boldsymbol{x}_i^{(t,m)}$ is the ith feature vector for task t and modality m, the partial derivative of this inner maximization problem is:

$$\frac{\partial O_{\boldsymbol{\eta}}}{\partial \boldsymbol{\eta}_m^{(t)}} = -2 \frac{\partial \Omega(\boldsymbol{\eta}^{(t)})}{\partial \boldsymbol{\eta}_m^{(t)}} - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \boldsymbol{\alpha}_i^{(t)} \boldsymbol{\alpha}_j^{(t)} y_i^{(t)} y_j^{(t)} \left(k_m^{(t)}(\boldsymbol{x}_i^{(t,m)}, \boldsymbol{x}_j^{(t,m)}) + \frac{\delta_{i,j}}{2C} \right)$$

We note that the gradient of the $L1$ regularizer is

$$\frac{\partial \Omega_1(\boldsymbol{\eta}^{(t)})}{\partial \boldsymbol{\eta}_m^{(t)}} = -\nu \sum_{s=1}^T \boldsymbol{\eta}_m^{(s)}$$

and the gradient of the $L2$ regularizer is

$$\frac{\partial \Omega_2(\boldsymbol{\eta}^{(t)})}{\partial \boldsymbol{\eta}_m^{(t)}} = -\nu \sum_{s=1}^T 2(\boldsymbol{\eta}_m^{(t)} - \boldsymbol{\eta}_m^{(s)})$$

For our implementation of this algorithm, we determined that the $\boldsymbol{\eta}$'s had converged when the maximum change in any single task's kernel weights ($\boldsymbol{\eta}^{(t)}$) was less than $\epsilon = 0.001$. In addition, we used a fixed step size of 0.01. This could be improved in future work by using a line search.

Algorithm 1 MTMKL Algorithm

1: Initialize $\boldsymbol{\eta}^{(t)}$ as $(1/T, ..., 1/T)$, $\forall t$
2: **while** not converged **do**
3:     Solve each single-task, multi-kernel support vector machine using $\boldsymbol{\eta}^{(t)}$, $\forall t$
4:     Update $\boldsymbol{\eta}^{(t)}$ in the direction of $-\partial O_{\boldsymbol{\eta}}/\partial \boldsymbol{\eta}^{(t)}$, $\forall t$
5: **end while**

4.3 Validation and testing

To choose the optimal parameter settings for each algorithm, we iterate over many possible combinations of the ν and C parameters, the regularization functions ($L1$ and $L2$), and the kernel function and kernel parameters. After finding the optimal parameters, we test the classifier by obtaining five bootstrapped samples of the held-out test data, testing on each, and averaging the results.

5 Results

To assess the usefulness of MTMKL, we compare it to both a standard SVM and MKL. We found that the optimal parameter settings were as follows: SVM used a linear kernel with $C = 0.1$, MKL used an Radial Basis Function (RBF) kernel with $\beta = 0.01$ and $C = 100$, and MTMKL used an RBF kernel with $\beta = 0.01$, $C = 10$, $\nu = 0.01$, and an $L2$ regularizer.

Table 1: Results for classifying the five wellbeing measures on held-out test data

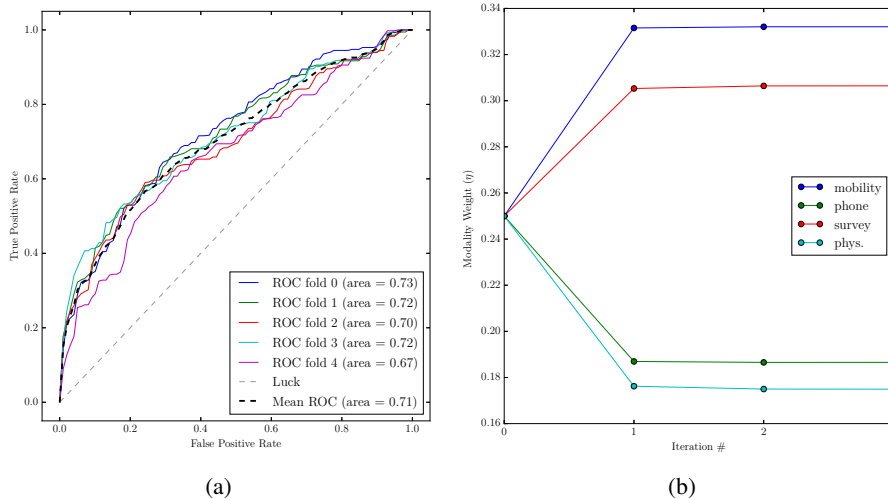|	SVM		MKL		MTMKL	
Task	**Accuracy**	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**
Happiness	68.48%	0.7535	69.58%	0.7593	74.30%	0.7963
Health	69.57%	0.7662	60.64%	0.7128	68.82%	0.7332
Alertness	67.08%	0.6748	59.48%	0.5453	56.41%	0.6029
Energy	60.25%	0.6102	56.66%	0.5331	64.40%	0.7087
Stress	69.57%	0.7981	66.45%	0.7106	70.91%	0.7465
All	-	-	-	-	64.80%	0.7050

(a)   (b)

Figure 1: (a) ROC Curve for the MTMKL Classifier on the 5 bootstrapped samples of the held-out test data (b) Kernel weight (η) convergence in happiness task, the other four tasks had very similar convergence paths and final η values

Table 1 shows the performance of the three algorithms in classifying each component of wellbeing. Surprisingly, we see that MKL generally underperforms a simple SVM; evidently, the implementation of MKL requires further improvement. In spite of this, MTMKL offers performance markedly exceeding that of both SVM and MKL in the target task of classifying happiness, suggesting there is an added benefit of learning the tasks together. For each additional task (health, alertness, energy, and stress), the SVM and MKL models were re-trained and the parameters were tuned to optimize performance on that task specifically. In contrast, the parameters for the MTMKL classifier were chosen to optimize for performance in classifying happiness. In spite of this, MTMKL still offers similar or superior performance in four of the five tasks. However, performance in classifying alertness is low. Upon further investigation we found that optimizing for alertness increased MTMKL's performance in classifying health, stress, and alertness, although accuracy in classifying alertness still did not exceed that of SVM. We will investigate how to further improve MTMKL's performance on all tasks in future work. For now, we emphasize that MTMKL provides the ability to classify each of the wellbeing dimensions simultaneously, after training and optimizing only one model. Since the ultimate goal of this research is to be able to predict a student's overall wellbeing, this is a compelling benefit.

Figure 1a shows the Receiver Operating Characteristic (ROC) curves obtained from classifying overall wellbeing using MTMKL on each bootstrapped sample of the held-out test data. Figure 1b shows how the η values for the happiness task change as the algorithm is trained. We can see that very few iterations are required before the optimal values are found. Note that the strongest weight is placed on the Mobility modality (confirming previous results that Mobility led to the highest classification accuracy for happiness [7]), while the weights for the Physiology and Phone modalitues are relatively low. The other four tasks showed the same pattern. This is an interesting result; if MTMKL could automatically determine that a given data modality provides little additional benefit

over the others, then it might be possible to reduce the overhead of running the study by limiting the collection of this data type. However, these results are preliminary; the low weight could also indicate that the feature extraction process for this modality requires improvement, or that more data is required before such high-dimensional modalities can offer performance enhancements.

6 Conclusions and Future Work

This work has presented how MTMKL, a recently developed technique combining Multiple Kernel Learning (MKL) and Multi-task Learning (MTL), is well-suited to the problem of using multi-modal data to model complex internal states like wellbeing. MKL is used to learn the importance of each data modality, which could help researchers assess what types of data need to be gathered when running expensive user studies. MTL is used to simultaneously model several dimensions of wellbeing as related tasks, so that each task can benefit from the others. We show that MTMKL provides performance improvements over both a traditional SVM classifier and over MKL, and is able to classify each of the wellbeing dimensions within a single model.

This work is only a first step, and many improvements and extensions can be made. Most obviously, MTL can be applied so that each user is treated as a task, allowing the classifier to account for individual differences, which can be significant in ambiguous outcomes like wellbeing. Because the results revealed a strong benefit of multi-tasking in the problem, investigating other multi-task learning methods such as GOMTL [9] or MLMTL [13] could be beneficial. There are also several modifications that could be made to improve the MTMKL algorithm, such as using different kernel types for each modality, testing different convergence criteria, and using a line search to improve the gradient descent algorithm. Further, the dataset and features used were selected to facilitate comparison with previous work, and were not optimized for the MTMKL problem. To extend MTMKL, we could also define more fine-grained modalities (for example, skin temperature vs. skin conductance features), or potentially even define one kernel per feature to perform feature selection simultaneously with the classifier's optimization function.

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] W. Boucsein. *Electrodermal activity*. Springer Science+Business Media, LLC, 2012.

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

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

[5] Kyriacos Chrysostomou, M Lee, SY Chen, and X Liu. Wrapper feature selection., 2009.

[6] M. Gönen and E. Alpaydın. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268, 2011.

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

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

[9] Abhishek Kumar and Hal Daume III. Learning task grouping and overlap in multi-task learning. *arXiv preprint arXiv:1206.6417*, 2012.

[10] 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.

[11] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010.

[12] J. Ratey. *Spark: The revolutionary new science of exercise and the brain*. Hachette Digital, Inc., 2008.

[13] Bernardino Romera-Paredes, Hane Aung, Nadia Bianchi-Berthouze, and Massimiliano Pontil. Multilinear multitask learning. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1444–1452, 2013.

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

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

[16] A. Sano et al. Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones. In *Body Sensor Networks*, 2015.

[17] A. Sano and R. Picard. Stress recognition using wearable sensors and mobile phones. In *ACII*, pages 671–676. IEEE, 2013.

[18] M. Seligman. *Flourish: A visionary new understanding of happiness and well-being*. Simon and Schuster, 2012.

[19] N. Tsuno et al. Sleep and depression. *J. of Clin. Psychiatry*, 2005.

[20] J. Westefeld and S. Furr. Suicide and depression among college students. *Professional Psychology: Research and Practice*, 18(2):119, 1987.

[21] C. Xu et al. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.