Formula: FactORized MUlti-task LeArning for task discovery in personalized medical models

Jianpeng Xu[*] Jiayu Zhou[†] Pang-Ning Tan[*]

Abstract

Medical predictive modeling is a challenging problem due to the heterogeneous nature of the patients. In order to build effective medical predictive models we need to address such heterogeneous nature during modeling and allow patients to have their own personalized models instead of using a one-size-fits-all model. However, building a personalized model for each patient is computationally expensive and the over-parametrization of the model makes it susceptible to the model overfitting problem. To address these challenges, we propose a novel approach called FactORized MUlti-task LeArning model (Formula), which learns the personalized model of each patient via a sparse multi-task learning method. The personalized models are assumed to share a low-rank representation, known as the base models. Formula is designed to simultaneously learn the base models as well as the personalized model of each patient, where the latter is a linear combination of the base models. We have performed extensive experiments to evaluate the proposed approach on a real medical data set. The proposed approach delivered superior predictive performance while the personalized models offered many useful medical insights.

1 Introduction

Predictive modeling has become an integral component of many industries to deliver accurate predictions for various purposes, such as decision making and risk management. With the growing development and availability of electronic medical records (EMR), the practitioners in many clinical decision support and care management systems resort to leveraging patients' medical records to perform various predictive modeling tasks for risk predictions and disease analysis. Moreover, in many clinical and pharmaceutical researches, predictive models such as disease progression models are used to study the pathologies of disease and evaluate the effectiveness of treatments, given the historical observations and medical records [4]. In the study of Alzheimer's disease (AD), for example, various predictive models are designed to study the courses of the disease and its progression patterns, to identity sensitive biomarkers that signal progression of the disease, and to build accurate models that identify high risk patients [25, 26].

Compared to standard data mining and machine learning applications, medical predictive modeling is especially challenging due to the heterogeneous nature of the patients. The heterogeneity arises from multiple factors: first of all, although some patients have similar phenotypes according to their health records, their medical conditions may vary. For example, in the study of dementia, patients with similar cognitive impairments may have different pathological causes. Another example is the study of heart failure (HF), where HF may be caused by coronary artery disease, hypertension, impaired glucose tolerance, and other factors [12, 16]. Secondly, it is well acknowledged that patients with the same disease may progress differently [17]. As such, one should address the heterogeneity of the patients in order to build accurate medical predictive models. It is widely accepted that building personalized models [15] is key to solving the problem, taking the inherent variability of the patients into account.

One simple way to implement the personalized models is to build a separate model for each patient independently. However, there are several drawbacks of this 'fully personalized' approach: First, it is not efficient in terms of time and space complexity. The task of building the personalized models is expensive and storing them is infeasible when the number of patients is large. More importantly, this approach requires solving a predictive modeling problem with a huge number of parameters. Because we have only limited amount of training data, such models are likely to severely overfit the data and result in models with poor generalization performance.

Instead of building a different model for each patient, an alternative approach is to consider the similarity of the patients. Specifically, a two-stage modeling is performed—grouping the patients first based on their similarities and then building a separate model

[*] Computer Science and Engineering Department, Michigan State University

[†] Samsung Research America, San Jose, CA

for patients in each group independently. This includes methods such as locally weighted learning [2] and localized support vector machine (LSVM) [9]. Locally weighted learning is a lazy learning scheme, in which the learning procedure only starts when the testing is performed. This approach would find the neighbors of the test instance, forming a group centered at the instance. It then builds a predictive model based on the training instances in the group. LSVM [9] is another approach, where supervised clustering is initially performed to group the training instances. It then trains a local SVM model for each cluster independently. One potential limitation of the two-stage approach is that the training of a model for patients within each group does not utilize potentially valuable information about patients from other groups since the grouping and model building steps are carried out separately. In addition, the approach is not exactly personalized since all the patients that belong to the same group have the same predictive model.

To address these limitations, this paper introduces a novel approach called FactORized MUlti-task LeArning model (FORMULA). FORMULA learns a personalized model for each patient in a tractable way by assuming the models share a low-rank representation, known as its 'base models'. The personalized model for each patient is a linear combination of a few of these base models. The base models can be regarded as features characterizing the underlying groups of the data while the coefficients of these base models denote memberships of the patients in these groups. To ensure the robustness of FORMULA, we enforce sparsity in both the 'base models' as well as the combination coefficients. As a result, each base model involves only a few relevant features, while each personalized model is a linear combination of only a few base models. FORMULA also enforces a graph Laplacian regularization to ensure that the personalized models for similar patients should be close to each other.

In short, the main contributions of this paper are summarized below:

- We proposed a novel personalized medical model called FORMULA. Instead of building a single model for all the patients or applying a two-stage modeling, FORMULA extracts the base models of the patients and uses a linear combination of these models as the personalized model of a patient.

- We employed a sparse matrix factorization formulation to perform base model selection for each patient and feature selection for each base model.

- We designed an efficient optimization method to solve this non-convex problem.

- We evaluated FORMULA on the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The experimental results show the superiority of FORMULA over other baseline methods.

The remainder of this paper is organized as follows. A brief review of the related works is given in Section 2. In Section 3.1, we formalize the problem of learning personalized models. The proposed FORMULA approach is introduced in Section 3.2. We solve its corresponding optimization problem in Section 3.3. Section 4 evaluated the performance of FORMULA on a real world dataset. Finally, we conclude the paper in Section 5.

2 Related Works

As mentioned in Section 1, in order to avoid learning one model for each patient (data point), we might consider either locally weighted learning, or two-stage learning methods, such as clustering plus multi-task learning. In this section, we are going to review locally weighted learning and multi-task learning.

2.1 Locally Weighted Learning Locally weighted learning is categorized as lazy learning method [2], in which the model is learned only when the testing data point comes. Locally weighted learning has been imbedded into various kinds of fundamental approaches, such as locally weighted regression [2], localized SVM [9], etc. The drawback of locally weighted learning is that it needs to build one model for each testing data points [22]. To address this drawback, localized SVM proposed an efficient learning method by first clustering the training data points into different groups and then building an SVM for each group. However, localized SVM need to take into consideration all the testing data points in advance to do the clustering over the training data samples, but usually the testing data set might not be available in advance. Also it does not address the problem when a new testing point becomes available, whether the method will retrain all the models using all available testing points, or just use the already generated models for each group. As mentioned earlier, localized SVM is a two-stage model and it builds models independently between groups. FORMULA is different from locally weighted learning in that it benefits from learning the group and the models simultaneously, and also considers the relations between groups implicitly.

2.2 Multi-Task Learning (MTL) Multi-task learning [6] is a methodology designed to improve predictive performance that learns different tasks simultaneously by taking into consideration the relations between tasks. The key difference between various MTLs lies in the way how they define the task relations. For example, the task relationships can be modeled using a common prior within a hierarchical Bayesian framework [3, 21], or using different kinds of regularization techniques, such as Mean-regularized MTL [11], low-rank regularized MTL [8], MTL with joint feature learning [28, 23], etc. Multi-task learning has also been used for feature learning and selection [1] or temporal learning [28, 20] by considering each time point as one task. An open-source multi-task learning software package MALSAR [24] has been developed to include efficient solvers for many state-of-the-art multi-task learning algorithms. As the second stage in the two-stage modeling scheme, multi-task learning can explicitly consider the relations between different groups/tasks. Although they can utilize the relations between tasks, the relations are mostly predefined. If the predefined task relations do not reflect the true underline relation, the performance will be degenerated. In this paper, we can address this problem by incorporating task relations implicitly, and identify the tasks and learn the task models simultaneously.

3 Learning Personalized Model via FORMULA

In this section, we formally introduce the problem of learning personalized models. We then discuss the technical challenges of the problem, which motivate the proposed FORMULA approach. Finally, we discuss how the problem can be efficiently solved.

3.1 Problem Formulation In a typical predictive modeling setting, we are given a feature vector and a target variable for each data point. Our goal is to learn a model that predicts the value of the target variable given its feature vector. In the context of medical predictive modeling, the features can be extracted from various sources, including historical medical records or medical images. The target variable can be binary-valued, such as the onset of a certain disease, or continuous-valued, such as the cognitive score of a patient.

Let $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ denote a collection of N training samples, where each sample is characterized by a D-dimensional feature vector, $\mathbf{x}_i \in \mathcal{R}^D$, and a target response[1] $y_i \in \Re$. We assume that the target can be approximated by a linear combination of the features, i.e., $y_i = \mathbf{w}_i^T \mathbf{x}_i + \epsilon_i$, where $\mathbf{w}_i$ is the parameter vector associated with the i-th training sample and ϵ_i is the Gaussian noise term. Since we are seeking for personalized models, the parameters $\mathbf{w}_i$ are unique for each sample and are estimated by solving the following optimization problem:

$$(3.1) \qquad \min_W \sum_{i=1}^N \ell_i(\mathbf{x}_i, y_i; \mathbf{w}_i)$$

where $\ell_i(\mathbf{x}_i, y_i; \mathbf{w}_i)$ is the loss function for sample i. For brevity, we use the notation $W = [\mathbf{w}_1, ..., \mathbf{w}_N] \in \mathcal{R}^{D\times N}$ to denote the model matrix.

Since there are $D \times N$ parameters that must be estimated from the N training samples, this leads to an underdetermined system of linear equations, which has either no solution or infinitely many solutions. The over-parameterization of the model also makes it susceptible to model overfitting. To overcome these problems, the number of effective parameters must be significantly reduced. One way to achieve this is by identifying groups of similar samples and then build a separate model for each group. Let $\mathcal{G} = \{\pi_1, \pi_2, \cdots, \pi_K\}$ denote the set of K groups, where π_j denote the set of samples assigned to the j-th group. The problem of learning personalized models for each group can be formalized as follows:

$$(3.2) \qquad \min_{W,\mathcal{G}} \sum_{\pi_j \in \mathcal{G}} \sum_{(\mathbf{x}_i, y_i)\in\pi_j} \ell_i(\mathbf{x}_i, y_i; \mathbf{w}_j)$$

The optimization problem can be solved using a two-stage approach, where the group membership information is initially obtained by applying clustering techniques such as k-means. Once the clusters are found, a personalized model is derived for each cluster by solving the inner summation term of the objective function given in (3.2). However, since the clustering is performed independently of the predictive modeling step, this may lead to suboptimal performance as the construction of the model for each group does not utilize information from other groups. The multi-task learning approach to be described in the next section is designed to overcome this problem by solving the clustering and predictive modeling steps jointly in a unified learning framework, thus supporting knowledge transfer among the clusters. In addition, to improve robustness of the predictions, additional sparsity constraints were imposed to further reduce the number of effective parameters that the models depend upon.

3.2 The Proposed FORMULA Framework This section presents the proposed FORMULA approach, which considers the development of personalized model for each patient as a single learning task. Unlike the two-stage approach given in Equation (3.2), FORMULA assumes the learning tasks are related. It therefore simultaneously learns the related tasks and utilize the shared information among tasks to improve its overall predictions.

[1] In this paper, we focus on the regression problem.

We achieve these goals by incorporating regularization terms into the personalized model formulation given in (3.1):

$$\min_{W} \mathbf{L}(X, \mathbf{y}; W) + \mathbf{R}(W) \tag{3.3}$$

where $\mathbf{L}(X, \mathbf{y}; W) = \sum_{i=1}^{N} \ell_i(\mathbf{x}_i, y_i; \mathbf{w}_i)$ is the loss function and $\mathbf{R}(W)$ is the regularization term, which encodes our modeling assumptions. To start with, we consider the following modeling assumptions of our formulation:

- **Model Clustering.** One of the key assumptions behind our proposed approach is that the predictions of the target variables are governed by a set of K *base models*, which are collectively represented by the matrix $U \in \mathcal{R}^{D\times K} = [\mathbf{u}_1, \ldots, \mathbf{u}_K]$, where each base model is represented by a column vector $\mathbf{u}_i \in \mathcal{R}^D$. We further assume that each personalized model $\mathbf{w}_i$ is represented by a linear combinations of the base models, i.e., $\mathbf{w}_i = U\mathbf{v}_i = \sum_{j=1}^{K} \mathbf{u}_j v_{ij}$, where $\mathbf{v}_i \in \mathcal{R}^K$ is a vector denoting the coefficients of the linear combination, and $V \in \mathcal{R}^{K\times N} = [\mathbf{v}_1, \ldots, \mathbf{v}_N]$. This assumption can be enforced by requiring the model matrix W to be as close as possible to the product of two matrices, i.e., $W = UV$.

- **Sparse Personalized Models.** Depending on the nature of the data, the number of base models can be potentially large. However, the personalized model of each individual patient is assumed to be a linear combination of only a few base models. In other words, the number of non-zero elements in V should be as few as possible. This can be achieved by enforcing a sparse-inducing norm on the matrix V. In addition, to ensure interpretability of the cluster assignment, the elements in V should be non-negative.

- **Sparse Base Models.** Each base model should be characterized by only a few relevant features, to ensure the model is robust to noise. A sparse-inducing norm can be applied to U to obtain the sparse base models.

- **Local Smoothness and Recovery.** Although each patient has its own personalized model, we assume the models for patients with similar phenotypes should be close to one another. Such a *model smoothness* criterion is helpful to infer the personalized model of a test patient by assuming it is similar to the weighted average of the personalized models for its neighbors. This can be achieved by incorporating a graph Laplacian regularization term into the proposed formulation.

Based on the preceding assumptions, the objective function of FORMULA is given by:

$$\begin{aligned} \min_{W,U,V} \; & \frac{1}{2}\sum\nolimits_{i=1}^{N} \ell_i(\mathbf{x}_i, y_i; \mathbf{w}_i) + \lambda_1\|V\|_1 + \lambda_2\|U\|_1 \\ & + \frac{\lambda_3}{2}\|W - WL\|_F^2 \\ \text{s.t. } & V \succeq 0, \; W = UV \end{aligned} \tag{3.4}$$

where $V \succeq 0$ denote all elements in V must be non negative and $L \in \mathcal{R}^{N\times N}$ is the similarity matrix between the training instances. The parameters λ_1, λ_2, and λ_3 control the tradeoffs among the various terms of the objective function. The last term in the objective function, $\|W - WL\|_F^2$, enforces the local smoothness constraint on the $\mathbf{w}_i$s. Note that L must be normalized such that the sum of each row or column is equal to 1. The number of base models K is assumed to be predefined by the user.

3.3 Optimization

This section describes how to solve the optimization problem for our proposed framework. In this work, we consider a squared loss function for regression problems, i.e., $\ell_i(\mathbf{x}_i, y_i; \mathbf{w}_i) = (y_i - \mathbf{w}_i^T\mathbf{x}_i)^2$. However, the optimization strategies used in this paper can also be applied to other loss functions. The objective function for FORMULA with squared loss is given by:

$$\begin{aligned} \min_{W,U,V} \; & \frac{1}{2}\sum\nolimits_{i}^{N} (y_i - \mathbf{w}_i^T\mathbf{x}_i)^2 + \lambda_1\|V\|_1 \\ & + \lambda_2\|U\|_1 + \frac{\lambda_3}{2}\|W - WL\|_F^2 \\ \text{s.t. } & V \succeq 0, W = UV \end{aligned}$$

We can simplify the problem by replacing W with the matrix product UV in the objective function, i.e., $\mathbf{w}_i = U\mathbf{v}_i$. This reduces the objective function to the following expression:

$$\begin{aligned} \min_{U,V} \; & \frac{1}{2}\sum\nolimits_{i}^{N} (y_i - \mathbf{v}_i^T U^T\mathbf{x}_i)^2 + \lambda_1\|V\|_1 \\ & + \lambda_2\|U\|_1 + \frac{\lambda_3}{2}\|UV - UVL\|_F^2 \\ \text{s.t. } & V \succeq 0 \end{aligned} \tag{3.5}$$

Thus, we only need to solve for U and V, and do not need to store the $D \times N$ matrix W. Similar to [7, 27], we propose to use the Block Coordinate Descent (BCD) algorithm to obtain a locally optimal solution. Specifically, we iteratively solve for U and V by fixing one of them to be constant, until the algorithm converges. Below we explain how each step can be solved efficiently.

Solve U, given V. The objective function becomes

$$\min_U \sum\nolimits_i^N (y_i - \mathbf{v}_i^T U^T \mathbf{x}_i)^2 + \frac{\lambda_2}{2}\|U\|_1 + \frac{\lambda_3}{2}\|UA\|_F^2$$

where $A = V(I - L)$. This is an ℓ_1-regularized convex optimization problem, which can be efficiently solved using projected gradient methods, such as spectral projected gradient[19], by considering the gradient of the smooth part of the objective function. Here, the gradient of the smooth part w.r.t. U is given by,

$$\sum\nolimits_{i=1}^N \left(-y_i \mathbf{x}_i \mathbf{v}_i^T + \mathbf{x}_i \mathbf{x}_i^T U \mathbf{v}_i \mathbf{v}_i^T\right) + \lambda_3 U A A^T$$

Solve V, given U. The objective function becomes

$$\min_V \frac{1}{2} \sum\nolimits_i^N (y_i - \mathbf{v}_i^T \tilde{\mathbf{x}}_i)^2 + \lambda_1 \|V\|_1 + \frac{\lambda_3}{2}\|UVB\|_F^2$$

where $\tilde{\mathbf{x}}_i = U^T \mathbf{x}_i$ and $B = I - L$. The problem can be solved in a similar way. The gradient of the smooth part of the objective function w.r.t. V is given by,

$$P + Q + \lambda_3 U^T U V B B^T \tag{3.6}$$

where $P_{i,j} = -y_i \tilde{\mathbf{x}}_{i,j}$, or $P_{i\cdot} = -y_i \tilde{\mathbf{x}}_i$; $Q_{i\cdot} = \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T \mathbf{v}_i$, and $\mathbf{v}_i$ is the i-th column of V.

4 Experimental Evaluation and Results

We have performed extensive experiments to evaluate the performance of Formula.

4.1 Dataset Our experiments were performed on the ADNI dataset[2], which contains images from MRI scans (M) and PET scans (P), as well as CSF measurements (C) and cognition-related clinical measurements such as Mini Mental State Examination (MMSE) scores and Alzheimer's Disease Assessment Scale-cognitive subscores (ADAS-Cog). ADNI is a longitudinal project, in which the measurements are collected repeatedly over a 6-month or 1-year interval. We call the time point when the patient came to the hospital for screening as *baseline*. The time point when the patient came to the hospital for evaluation is determined based on the elapsed time since the initial *baseline*. For example, M06 denote the time point 6 months after the first visit. There are altogether 5 time points, designated as M06, M12, M24, M36 and M48, respectively. We consider the samples collected for each time point as a separate data set. The sample sizes for the five data sets are shown in Table 1. Note that the data sets decrease in size due to the drop out of some patients for various reasons.

Table 1: Dataset size of ADAS-Cog and MMSE

	M06	M12	M24	M36	M48
ADAS-Cog	648	638	564	377	85
MMSE	648	642	569	389	87

The features of each data set include those from M, P, C and META (E), which denote additional features other than M, P and C. The detailed list of the META features is given in [28]. We consider using these features to build models for predicting the ADAS cognitive scores or MMSE scores on each data set.

4.2 Baseline Algorithms We compared the performance of Formula against the following baseline methods.

- Single model(SM): This is a one-size-fits-all approach, assuming there is no inherent groupings in the data. We applied ridge regression to construct a single model for each data set.
- Clustering + single task model with Ridge regression(CSTR): In this baseline algorithm, we first apply k-means clustering to generate k clusters. We then build a ridge regression model for each cluster.
- Clustering + single task model with Lasso regression(CSTL): This approach is similar to CSTR except we use Lasso regression to build the model instead of ridge regression.
- Clustering + sparse low rank mutli-task learning (CSL) [8]: First we cluster the data using k-means to generate k clusters, and then treat each cluster as a task to learn a multi-task model. Here we assume that all models share a low-rank representation in addition to a sparse property. The objective function for CSL is given by [8],

  $$\min_W \sum\nolimits_{i=1}^k \|X_i \mathbf{w}_i - \mathbf{y}_i\|_2^2 + \gamma \|P\|_1$$
  $$\text{s.t. } W = P + Q, \|Q\|_* \leq \tau$$

  where $W \in \mathcal{R}^{D \times K}$ and $W = [\mathbf{w}_1, ..., \mathbf{w}_k]$. We use the implementation in MALSAR [24]to solve CSL.
- Clustering + mean regularized multi-task learning (CMR) [11]: In this baseline algorithm, we first cluster the data using k-means into k clusters which are considered as k tasks. We consider a special task relation for the multi-task learning. The assumption is that the models for all tasks are close to their mean model. The objective function is

[2] Available at http://adni.loni.ucla.edu

given below.

$$\min_{W} \sum_{i=1}^{K} \|X_i\mathbf{w}_i - \mathbf{y}_i\|_2^2 + \rho_1 \sum_{i=1}^{K} \|\mathbf{w}_i\|_2^2 + \rho_2 \sum_{i=1}^{K} \left\|\mathbf{w}_i - \frac{1}{K} \sum_{s=1}^{K} \mathbf{w}_s\right\|_2^2$$

We use the implementation in MALSAR [24]to solve CMR.

4.3 Evaluation Method Each data set is partitioned into a training set, which contains 80% of the data points, and a test set, which contains the rest of the data. For SM, because there is only one model for the entire data set, we can apply the derived model directly on the test set for evaluation. For the two-stage methods (CSTR, CSTL, CSL and CMR), when the test data point becomes available, we first find the nearest cluster of each test point and apply its corresponding model to make the prediction. As previously mentioned, in FORMULA, the personalized model W are estimated for the training data only. For testing, we use the weighted average model of the nearest neighbors of each test point. Formally, the personalized model for the test point $\mathbf{x_i}$ is:

$$\mathbf{w_i} = \sum_{j=1}^{T} \frac{s_{i,j}}{\sum_{n=1}^{T} s_{i,n}} \mathbf{w}_j,$$

where T is the number of nearest neighbors of $\mathbf{x}_i$. We set $T = 5$ in our experiments. $s_{i,j}$ is the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. Here, $s_{i,j}$ is calculated using the Gaussian radial basis function. The model performance is evaluated based on its mean of squared error (MSE).

4.4 Experimental Results The experimental results are summarized in Table 2. The results suggest that FORMULA outperforms other baselines in 6 out of 10 data sets and is consistently one of the top two algorithms for all the data sets. In the 6 data sets in which FORMULA has the lowest MSE, its average improvement over the second best performer is more than 6%. In ADAS-Cog M36 and MMSE M12, the performance of FORMULA is almost the same as the top performers. Comparing FORMULA against CSL and CMR, even though all three methods are multi-task learning algorithms, FORMULA outperforms the other two consistently on all the data sets. The reason is that by learning the clusters and models simultaneously, FORMULA can utilize more information in the learning process. Comparing CSTR/CSTL against CSL/CMR, observe that their performances are quite similar. However, there are several cases where CSTL outperforms CSL/CMR, which might be due to the inaccurate assumption of the task relations in the multi-task learning formulation.

In addition, it is worth noting that the performance for all the methods degrade from M06 to M48. This is because the sample size decreases over time, which makes the methods more susceptible to underfitting their models with inadequate data points.

4.5 Sensitivity Analysis Since there are four parameters that must be tuned, namely, λ_1, λ_2, λ_3, and K, we need to analyze the sensitivity of the models to changes in the parameter values. By varying the value of one parameter and keeping the other three parameters constant, the sensitivity analysis results for λ_1, λ_2, λ_3, and K are shown in Figure 1. From Figure 1, we can see that the model is quite robust to changes in these parameter values. Nonetheless, in general, λ_1 prefers larger values (see Figure 1a) whereas λ_2 prefers smaller ones (see Figure 1b). The model is also not that sensitive to changes in λ_3 and K (Figure 1c, 1d) within the range of parameter values investigated.

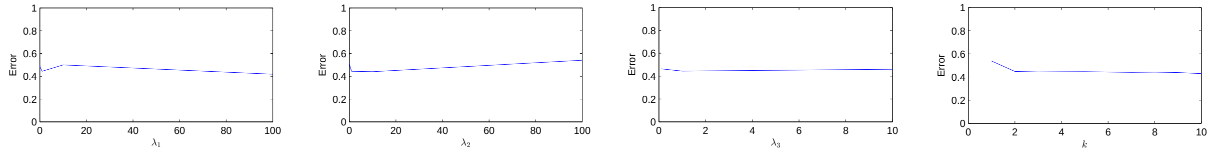
4.6 Model Analysis One of the most attractive feature of the proposed model is that it learns a set of base models from the training data, where each base model corresponds to a column of the matrix U. The personalized models can be represented by a linear combination of these base model. In this section, we first investigate the base models obtained from the best solution in the last section and then analyze how each of the base models contribute to the personalized models.

Base Models. In both ADAS-Cog and MMSE datasets, the rank of the best models obtained is 3. We sort the features in each base model and rank them according to their contribution. The top features for each base model in ADAS-Cog and MMSE are shown in Table 3 and Table 4, respectively. We observe that the top features in the base models look very different from each other. Due to the progression of Alzheimer's disease, it will eventually affect almost all parts of the brain. One possible explanation of these heterogeneous models is that, the patients may be at different stages of the disease, and at different stages, the contribution from the base model differs.

In the *Base Model A* of the ADAS-Cog task (Table 3), the leading feature is the cortical thickness average of the right lateral Occipital. The relationship between occipital and AD was studied by previous works [18, 5], and found to be significant in the advanced AD patients. In both *Base Model B* and *Base Model C*, the leading feature is the cortical thickness average of the left middle temporal gyrus, which is an important area in the Temporal lobe, connected to multiple cognitive functions, such as accessing word meaning while reading. The area is found to be the first temporal lobe

Table 2: Comparison the MSE between FORMULA and baseline methods

MSE	ADAS-Cog					MMSE				
	M06	M12	M24	M36	M48	M06	M12	M24	M36	M48
SM	0.497	0.553	**0.758**	**0.893**	2.153	0.237	0.331	0.420	0.536	1.202
CSTR	0.469	0.637	0.879	0.985	1.543	0.246	0.329	0.392	0.493	0.747
CSTL	0.448	0.558	0.817	0.907	**1.416**	0.205	**0.264**	0.287	0.364	0.954
CSL	0.555	0.695	0.869	0.981	1.524	0.247	0.335	0.383	0.470	0.821
CMR	0.545	0.625	0.860	0.978	1.524	0.242	0.319	0.362	0.420	0.786
FORMULA	**0.424**	**0.545**	0.764	0.899	1.520	**0.197**	0.267	**0.262**	**0.347**	**0.673**



(a) Effect of varying λ_1 on MSE (b) Effect of varying λ_2 on MSE (c) Effect of varying λ_3 on MSE (d) Effect of varying k on MSE

Figure 1: Sensitivity Analysis of FORMULA. Test is performed on ADAS-Cog M06 dataset.

neocortical sites affected in AD [10]. As such, these two models may relate to early predictive patterns of AD. We also find that in *Base Model B*, the effects of *Left Entorhinal* is higher than *White matter volumn of the left Hippocampus*, while in the *Base Model C* the pattern is reversed. The two models may indicate different progression patterns in early stage AD patients.

In the base models obtained in MMSE task (Table 4), we find that the leading features have different patterns. In *Base Model A*, the leading feature is the *average cortical thickness of the inferior parietal*. The area of inferior parietal is related to the progression of AD in several studies. In [14], the authors find its metabolism decreased early in the course of AD. The leading feature in *Base Model B* is the cortical parcellation volume of *right precentral*. The precentral gyrus is found to be significant in voxel analysis [13]. In *Base Model C*, the volume of *left hippocampus* dominates other features. This feature is considered as one of the most significant biomarkers of AD.

Contribution to Personalized Models. As we have learned heterogeneous base models for both tasks, it is interesting to see how each of the models contribute to the personalized models. Since both models are at rank 3, we are able to plot the coefficients in V using a 3-dimensional coordinator system. For each point, the value at an axis means how much the corresponding base model contributes to its personalized model. The scatter plots of the contributions are given in Figure 2. We are able to find very interesting patterns in these plots.

In Figure 2.a, the size of the marker is propotional to the value of ADAS-Cog score. Note that lower ADAS-Cog values indicate better cognitive functionality, i.e., cognitive normal patients have smaller markers in the plot. First of all, we see that only a few patients have (7 patients with only base model A, 31 patients with B, and 13 patients with C), and it is not hard to find out that these patients are characterized by high ADAS-Cog scores. We are able to see boundaries among different groups of patients (patient with only one base

Table 3: Base Models for the ADAS-Cog task

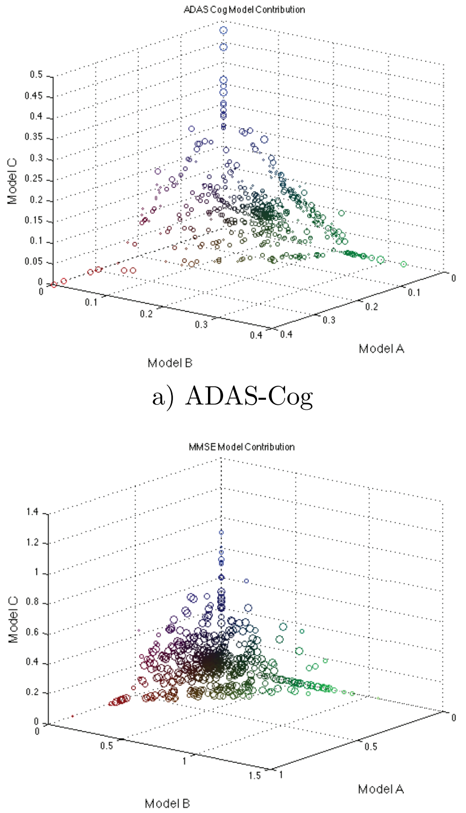
Base Model A	
CTA:R.Lateral Occipital	1.366
V-CP:R.Caudal Middle Frontal	1.287
SA:L.Middle Temporal	1.272
CTA:L.Middle Temporal	1.122
CTA:L.Rostral Middle Frontal	1.104
CTA:R.ParsTriangularis	0.966
Base Model B	
CTA:L.Middle Temporal	2.467
CTA:R.Rostral Middle Frontal	1.907
CTA:L.Entorhinal	1.726
V-WM:L.Hippocampus	1.523
SA:L.Middle Temporal	1.193
V-CP:R.Caudal Middle Frontal	1.168
Base Model C	
CTA:L.Middle Temporal	2.042
V-WM:L.Hippocampus	1.949
SA:L.Middle Temporal	1.843
CTA:L.Entorhinal	1.722
CTA:R.Rostral Middle Frontal	1.587
V-CP:R.Sup.Temporal	1.323

Table 4: Base Models for MMSE task

Base Model A	
CTA:L.Inf.Parietal	0.210
CTA:L.Middle Temporal	0.175
CTA:L.Lateral Occipital	0.167
CTA:L.Inf.Temporal	0.163
CTA:L.Sup.Parietal	0.159
V-WM:L.Hippocampus	0.153
Base Model B	
V-CP:R.Precentral	0.149
V-CP:L.Sup.Frontal	0.147
V-CP:R.Tra.Temporal	0.137
V-CP:R.Lingual	0.136
V-CP:R.Sup.Frontal	0.128
V-CP:L.Inf.Temporal	0.120
Base Model C	
V-WM:L.Hippocampus	0.216
SA:L.Pericalcarine	0.169
SA:R.Rostral Middle Frontal	0.167
V-CP:R.Rostral Middle Frontal	0.164
SA:L.Hemisphere	0.159
SA:L.ParsTriangularis	0.151

model, patients with two base models, and those with three base models). This is probably because that due to the ℓ_1 thrinkage effects, small contributions turn to zeros. We are able to see much more patients with linear combinations of base model B and C (112 patients in total), as compared to other two groups (39 patients for A and B, 36 patients for A and C). And we also notice the personalize models for most patients (280 patients in total) are linear combination of three base model.

The results for MMSE M06 task is given in Figure 2.b, in which the size of the marker is propotional to the value of the patient's MMSE score. Cognitive normal patients usually have higher MMSE scores, which means smaller markers indicates patients affected more by Alzheimer's. We are able to see that the personalized models for the largest population are linear combination of the three models (230 patients). Only a few patients lie on the axises (16 patients for base model A, 32 for B, and 27 for C) and those patients usually have smaller MMSE values as compared to the rest of population. Considering the difference between ADAS-Cog and MMSE, i.e., healthy patients typically have low ADAS-Cog score and high MMSE score, the finding in Figure 2.a and Figure 2.b are consistent: the models for patients with advanced Alzheimer's are more likely to be singleton (and heterogeneous). Also, the findings in this paper are consistent with our assumption that predictive models of the patients are not homogeneous, and for different set of patients, the models should be different.



a) ADAS-Cog

b) MMSE

Figure 2: Model contributions for personalized models.

There are many other interesting findings we have, for example, the patterns of different latent modality over the course of progression of Alzheimer's. We leave a complete analysis to future publication.

5 Conclusion

Personalized modeling for medical use is one of the emerging research topics in machine learning and data mining area, and there are many challenges associated with it. To address these challenges, we propose a novel FactORized MUlti-task LeArning model (Formula) to learn low-rank personalized models, leveraging the shared information among patients. Specifically, the proposed approach learns a personalized model for each patient, assuming the models share a low-rank representation. The personalized models are computed as linear combinations of a few base models. Our experimental results on the Alzheimer's Disease Neuroimaging Initiative (ADNI) data set suggest that the proposed approach is superior than several baseline methods and provide many valuable medical insights.

6 Acknowledgments

Jianpeng Xu and Pang-Ning Tan's research are partially supported by NOAA Climate Program office through grant NA12OAR4310081 and NASA Terrestrial Hydrology Program through grant NNX13AI44G.

References

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Mach. Learn.*, 73(3):243–272, Dec. 2008.

[2] C. G. Atkeson, A. W. Moore, and S. Schaal. Locally weighted learning. *JAIR*, 11(1-5):11–73, 1997.

[3] B. Bakker and T. Heskes. Task clustering and gating for bayesian multitask learning. *JLMR*, 4:83–99, Dec. 2003.

[4] R. Bellazzi, F. Ferrazzi, and L. Sacchi. Predictive data mining in clinical medicine: a focus on selected methods and applications. *DMKD*, 1(5):416–430, 2011.

[5] H. Braak, E. Braak, and P. Kalus. Alzheimer's disease: areal and laminar pathology in the occipital isocortex. *Acta Neuropathologica*, 77(5):494–506, 1989.

[6] R. Caruana. Multitask learning. *Mach. Learn.*, 28(1):41–75, 1997.

[7] S. Chang, G.-J. Qi, C. Aggarwal, J. Zhou, M. Wang, and T. Huang. Factorized similarity learning in networks. In *ICDM*, pages 917–926, 2014.

[8] J. Chen, J. Liu, and J. Ye. Learning incoherent sparse and low-rank patterns from multiple tasks. *TKDD*, 5(4):22:1–22:31, 2012.

[9] H. Cheng, P.-N. Tan, and R. Jin. Efficient algorithm for localized support vector machine. *TKDE*, 22(4):537–549, 2010.

[10] A. Convit, J. De Asis, M. De Leon, C. Tarshish, S. De Santi, and H. Rusinek. Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to alzheimer's disease. *Neurobiology of Aging*, 21(1):19–26, 2000.

[11] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *SIGKDD*, pages 109–117, 2004.

[12] K. Fox. Efficacy of perindopril in reduction of cardiovascular events among patients with stable coronary artery disease: randomised, double-blind, placebo-controlled, multicentre trial (the europa study). *Lancet*, 362(9386):782–788, 2003.

[13] G. Frisoni, C. Testa, A. Zorzan, F. Sabattoli, A. Beltramello, H. Soininen, and M. Laakso. Detection of grey matter loss in mild alzheimer's disease with voxel based morphometry. *Journal of Neurology, Neurosurgery & Psychiatry*, 73(6):657–664, 2002.

[14] M. D. Greicius, G. Srivastava, A. L. Reiss, and V. Menon. Default-mode network activity distinguishes alzheimer's disease from healthy aging: evidence from functional mri. *Proceedings of the National Academy of Sciences of the United States of America*, 101(13):4637–4642, 2004.

[15] M. A. Hamburg and F. S. Collins. The Path to Personalized Medicine. *New England Journal of Medicine*, 363(4):301–304, 2010.

[16] W. B. Kannel and A. J. Belanger. Epidemiology of heart failure. *American Heart Journal*, 121(3):951–957, 1991.

[17] S. Oddo, A. Caccamo, J. D. Shepherd, M. P. Murphy, T. E. Golde, R. Kayed, R. Metherate, M. P. Mattson, Y. Akbari, and F. M. LaFerla. Triple-transgenic model of alzheimer's disease with plaques and tangles: intracellular aβ and synaptic dysfunction. *Neuron*, 39(3):409–421, 2003.

[18] M. Penttilä, J. V. Partanen, H. Soininen, and P. Riekkinen. Quantitative analysis of occipital eeg in different stages of alzheimer's disease. *Electroencephalogr Clin Neurophysiol.*, 60(1):1–6, 1985.

[19] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. *Trans. Sig. Proc.*, 57(7):2479–2493, 2009.

[20] J. Xu, P.-N. Tan, and L. Luo. ORION: Online Regularized multI-task regressiON and its application to ensemble forecasting. In *ICDM*, pages 1061–1066, 2014.

[21] K. Yu, V. Tresp, and A. Schwaighofer. Learning gaussian processes from multiple tasks. In *ICML*, pages 1012–1019, 2005.

[22] H. Zhang, A. Berg, M. Maire, and J. Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *CVPR*, volume 2, pages 2126–2136, 2006.

[23] J. Zhou, J. Chen, and J. Ye. Clustered multi-task learning via alternating structure optimization. In *NIPS*, pages 702–710, 2011.

[24] J. Zhou, J. Chen, and J. Ye. *MALSAR: Multi-tAsk Learning via StructurAl Regularization*. Arizona State University, 2011.

[25] J. Zhou, J. Liu, V. A. Narayan, and J. Ye. Modeling disease progression via multi-task learning. *NeuroImage*, 78:233–248, 2013.

[26] J. Zhou, Z. Lu, J. Sun, L. Yuan, F. Wang, and J. Ye. Feafiner: Biomarker identification from medical data through feature generalization and selection. In *SIGKDD*, pages 1034–1042, 2013.

[27] J. Zhou, F. Wang, J. Hu, and J. Ye. From micro to macro: Data driven phenotyping by densification of longitudinal electronic medical records. In *SIGKDD*, pages 135–144, 2014.

[28] J. Zhou, L. Yuan, J. Liu, and J. Ye. A multi-task learning formulation for predicting disease progression. In *SIGKDD*, pages 814–822, 2011.