2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)
4-7 March 2018
Las Vegas, Nevada, USA

Obesity Risk Factors Ranking Using Multi-Task Learning

Lu Wang[1], Dongxiao Zhu[1]* , Elizabeth Towner [2] and Ming Dong[1]

***Abstract*— Obesity is one of the leading preventable causes of death in the United States (U.S.). Risk factor analysis is a process to identify and understand the risk factors contributing to a particular disease, and is an imperative component in the development of efficient and effective prevention and intervention efforts. Most existing methods usually aim to build a one-size-fits-all model to identify the risk factors at the population-level. However, this type of methods does not take into consideration of heterogeneity in the population. To overcome this limitation, we formulate the subpopulation specific obesity risk factors ranking problem, under the framework of multi-task learning (MTL), to identify a ranked list of obesity risk factors for each subpopulation (task) simultaneously with utilizing appropriate shared information across tasks. By synchronously learning multiple related tasks, MTL provides a paradigm to rank risk factors both at the subpopulation and population-levels.**

I. Introduction

At present, more than one-third (36.5%[1]) of adults living in the United States (U.S.) are obese. Obesity is one of the most common health threats and increases risk for negative health comorbidities, e.g., diabetes, metabolic syndrome and cardiovascular disease [1] and mortality [2]. Risk factor analysis has been extensively applied to identify, rank and understand the underlying factors for prevention and treatment of obesity, e.g., [3].

Risk factor analysis is a statistical method to learn the complex relationship between the dependent variable (i.e., target, outcome, output variable) and the independent variables (i.e., predictor or input variables), which is applied for the prevention, intervention and treatment of preventable diseases (e.g., obesity, cardiovascular disease, type 2 diabetes) [4], [5]. Conventional risk factor analysis utilizes either regression methods to estimate the relations between the dependent and independent variables, e.g., linear regression [6] and multivariate logistic regression [4], or standard statistical tests to distinguish the significant/influential factors using p-value, e.g., chi-square test [3], [7] and t-test [8].

The aforementioned risk factor analysis methods merely study the risk factors at the population-level, considered as single-task learning (STL) methods, which train a model for the entire population. Therefore, they fail to capture the heterogeneity in the population. However, the causes of obesity are multi-faceted and include both subpopulation-level and population-level risk factors as well as obesity influences some subpopulations more than others [9]. Subpopulations, e.g., people in various ages, diverse races, different living regions, etc., can be vastly different in their risk factors for obesity; so that precise identification and ranking of shared and unique risk factors for specific subpopulations are necessary to maximize the efficiency and effectiveness of obesity prevention and intervention efforts.

To conduct the risk factor analysis at both subpopulation and population-levels, multi-task learning (MTL) is proposed to rank the obesity risk factors for each subpopulation by training multiple models simultaneously that incorporate the heterogeneity between subpopulations [10]. Also, these subpopulations may share some common information, which is considered as the homogeneity in the population. To take into account the homogeneity in the population, we hypothesize that the multiple tasks are related though sharing a common set of risk factors among tasks. To test this hypothesis, we implement MTL with $l_{2,1}-$norm regularization across all tasks with a joint sparsity, which means each feature weight is either small or large for all individuals [11].

In this paper, we conduct the obesity risk factors analysis to implement a precise prevention, intervention and treatment plan for both subpopulation and population simultaneously. Fig. 1 demonstrates our MTL idea and compares with STL approaches.

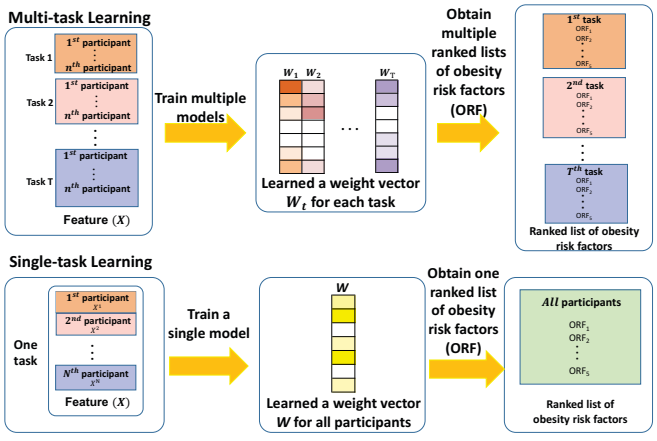

Fig. 1: **MTL trains multiple models simultaneously to obtain multiple ranked lists of obesity risk factors, i.e., one ranked list of obesity risk factors for each subpopulation, whereas STL trains a single one-size-fits-all model to obtain a ranked list of obesity risk factors for all subpopulations. Note that, in the learned weight vector, color box indicates a higher weight and white box means the weight is zero.**

The rest of this paper is organized as follows: Section II presents the MTL framework and the obesity risk factors ranking algorithm. In Section III, the effectiveness of the

*Corresponding Author
[1]Dept. of Computer Science, Wayne State University, Detroit, MI 48202, USA. {lu.wang3, dzhu, mdong}@wayne.edu
[2]Dept. of Family Medicine and Public Health Sciences, Wayne State University, Detroit, MI 48202, USA. ekuhl@med.wayne.edu
[1]https://www.cdc.gov/obesity/data/adult.html

MTL framework for ranking obesity risk factors is demonstrated using a public data set. Finally, in Section IV, we conclude with discussion and give future research directions.

II. Method

In this section, we start with the MTL framework and define notations, which are used throughout the paper. We then present the obesity risk factors ranking algorithm.

A. Multi-task learning framework

In the real-world scenario, multiple related tasks are more common than many independent tasks. Let us begin with the common object function for MTL that is used to minimize the penalized empirical loss. Assume there are T tasks and G continuous features in the data set, then we have the weight matrix as $W \in R^{T\times G}$ and MTL object function as:

$$\min_{W} \mathcal{L}(W) + \Omega(W), \tag{1}$$

where $\mathcal{L}(W)$ is the empirical loss function and $\Omega(W)$ is the regularization/penalty term, which encodes the task relatedness.

In MTL, $\mathcal{L}(W)$ can be expressed as:

$$\mathcal{L}(W) = \frac{1}{2}\sum_{t=1}^{T} \left\| X_t W_t^T - Y_t \right\|_F^2, \tag{2}$$

where t is the index of the task and $X_t \in \mathbb{R}^{n_t\times G}$ is the input matrix of the t^{th} task. Y_t is the corresponding target value, which is Body Mass Index (BMI) in our data set. $\|\cdot\|_F$ is Frobenius norm.

To make sure that the weight of each feature is either small for all participants or large for all participants, $l_{2,1}-$norm regularization:

$$\Omega(W) = \|W\|_{2,1}, \tag{3}$$

is used as the regularization term, where $\|W\|_{2,1} = \sum_{g=1}^{G}\sqrt{\sum_{t=1}^{T}|w_{tg}|^2}$. Note that, g is the index of feature, w_{tg} is the weight for the g^{th} feature in the t^{th} task.

Thus, we have the MTL object function as:

$$\min_{W} \frac{1}{2}\sum_{t=1}^{T} \left\| X_t W_t^T - Y_t \right\|_F^2 + \lambda \sum_{g=1}^{G}\sqrt{\sum_{t=1}^{T}|w_{tg}|^2}, \tag{4}$$

where $\lambda \geq 0$ is the tuning parameter that is used to control the relative impact of loss function and regularization term on the regression coefficient estimates. Larger value of λ produces more sparse weight matrix.

B. Obesity risk factors ranking algorithm

The optimization problem proposed in Eq. (4) is a standard $l_{2,1}$-norm regularization problem, which can be solved efficiently via the proximal gradient descent based algorithm that is summarized in Algorithm 1 as below.

Algorithm 1 outlines the learning procedure of proximal gradient descent algorithm to solve optimization problem in Eq. (4). In line 3, the search points in the i-th iteration ($S^{(i)}$) are an affine combination of $W^{(i)}$ and $W^{(i-1)}$. Line 6 is the building block for proximal gradient descent algorithm, where $\pi_P(\cdot)$ is the $l_{2,1}$-regularized Euclidean projection:

Algorithm 1: Proximal gradient descent algorithm for $l_{2,1}$-norm regularization problem.

Input: A set of feature matrices $\{X_1, X_2, \cdots, X_T\}$ and target value matrix Y for all T tasks, Initial coefficient matrix $W^{(0)}$, λ

Output: $\bar{W}$

1 **Initialize**: $W^{(1)} = W^{(0)}$, $d_{-1} = 0$, $d_0 = 1$,$\gamma_0 = 1$,$i = 1$;
2 **repeat**
3 Set $\alpha_i = \frac{d_{i-2}-1}{d_{i-1}}$, $S^{(i)} = W^{(i)} + \alpha_i(W^{(i)} - W^{(i-1)})$;
4 **for** $j = 2, 1, \cdots$ **do**
5 Set $\gamma = 2^j\gamma_{i-1}$;
6 Calculate $W^{(i+1)} = \pi_P(S^{(i)} - \frac{1}{\gamma}g^{'}(S^{(i)}))$;
7 Calculate $Q_\gamma(S^{(i)}, W^{(i+1)})$;
8 **if** $g(W^{(i+1)}) \leq Q_\gamma(S^{(i)}, W^{(i+1)})$ **then**
9 $\gamma_i = \gamma$, **break** ;
10 **end**
11 **end**
12 $d_i = \frac{1+\sqrt{1+4d_{i-1}^2}}{2}$;
13 $i = i + 1$;
14 **until** *Convergence of* $W^{(i)}$;
15 $\bar{W} = W^{(i)}$;

$$\pi_P(G(S^{(i)})) = \min \frac{1}{2} \parallel W - G(S^{(i)})) \parallel_F^2 + \lambda \parallel W \parallel_{2,1}, \tag{5}$$

where $G(S^{(i)}) = S^{(i)} - \frac{1}{\gamma_i}\Delta\mathcal{L}(S^{(i)}$ is a "gradient" step of $S^{(i)}$, and the gradient of the empirical loss function can be calculated as:

$$\Delta\mathcal{L}(S^{(i)}) = [(X_1 {S_1^{(i)}}^T - Y_1)X_1, ((X_2 {S_2^{(i)}}^T - Y_2)X_2, \cdots, ((X_T {S_T^{(i)}}^T - Y_t)X_T]. \tag{6}$$

An efficient solution (Theorem 1) of Eq. (5) has been proposed in [12].

Theorem 1: Given λ, the primal optimal point $\hat{W}$ of Eq.(5) can be calculated as:

$$\hat{W}_j = \begin{cases} \left(1 - \frac{\lambda}{\|G(S^{(i)})_j\|_2}\right) G(S^{(i)})_j & \text{if } \lambda > 0, \parallel G(S^{(i)})_j \parallel_2 > \lambda \\ 0 & \text{if } \lambda > 0, \parallel G(S^{(i)})_j \parallel_2 \leq \lambda \\ G(S^{(i)})_j & \text{if } \lambda = 0 \end{cases} \tag{7}$$

where $G(S^{(i)})_j$ is the j^{th} row of $G(S^{(i)})$, and $\hat{W}_j$ is the j^{th} row of $\hat{W}$.

In lines 4-11, the optimal γ_i is chosen by the backtracking rule based on [13, Lemma 2.1, page 189], γ_i is greater than or equal to the Lipschitz constant of $g(\cdot)$ at $S^{(i)}$, which means γ_i is satisfied for $S^{(i)}$ and $\frac{1}{\gamma_i}$ is the possible biggest step size.

In line 7, $Q_\gamma(S^{(i)}, W^{(i+1)})$ is the tangent line of $g(\cdot)$ at $S^{(i)}$, which can be calculated as:

$$Q_\gamma(S^{(i)}, W^{(i+1)}) = g(S^{(i)}) + \frac{\gamma}{2} \parallel W^{(i+1)} - S^{(i)} \parallel^2 + \langle W^{(i+1)} - S^{(i)}, g^{'}(S^{(i)}) \rangle.$$

III. Experiments and Results

To evaluate the performance of MTL for ranking obesity risk factors, we extensively compare it with STL methods. We first describe the setup of experiments and then describe the data set we use. Finally, we discuss the results on the ranked list of obesity risk factors obtained by MTL and STL methods.

A. Experiments setup

In the experiments, MTL with $l_{2,1}-$norm regularization is implemented in Matlab language [14], while two STL methods, i.e., linear model using generalized least squares (LMGLS) and linear mixed-effects model (LMEM), are implemented in R language using package *nlme* [15]. LMGLS is trained using the *gls* function in the *nlme* [15], which allows the errors to be correlated.

LMEM is extended from linear regression models for the data, which is collected and summarized in the grouping structure. LMEM is composed of two parts, i.e., fixed and random effects. Fixed-effects terms usually correspond to the traditional linear regression, while the random effects are associated with subpopulation experimental units that are drawn randomly from a population. The random effects have a distribution whereas fixed effects do not. The covariance structure is represented by LMEM and related to the data with grouping structure by associating the common random effects to observations, which have the same level of a grouping variable. We treat categorical features as factors with different levels, while consider continuous features as random effects. LMEM is trained using *lme* function in the *nlme* [15], which allows nested random effects and within-group errors to be correlated.

B. Behavioral risk factor surveillance system (BRFSS) data set

The BRFSS data set is a collaborative project between all the states in the U.S. and the Centers for Disease Control and Prevention (CDC), and aims to collect uniform, state-specific data on preventive health practices and risk behaviors that affect the health of the adult population (i.e., adults aged 18 years and older). In the experiments, we use the BRFSS data set that is collected in 2016[2]. The BRFSS data set is collected via the phone-based surveys with adults residing in private residence or college housing.

Considering that the sample size of a typical obesity study data is usually limited, we randomly sample a subset of data from the original BRFSS data set to validate our MTL method. We obtain the sample size $2,000$ based on the sample size estimation formula [16, Eq. 3.1, page 44].

[2]https://www.cdc.gov/brfss/annual_data/annual_2016.html

After deleting the entries with missing and hidden values, the preprocessed data set contains $2,000$ participants with 91 variables including 90 input variables and one output variable, i.e., BMI.

C. Experimental results and discussion

We define tasks based on various age groups and geographic information to generate two different MTL settings: A. Tasks in MTL1 are defined in terms of the predefined age groups by [17] and its results are shown in Fig. 2; B. Tasks in MTL2 are defined in terms of a geographic variable (i.e., states in the U.S.) and its results are shown in Fig. 3.

EnergyDaysPerMonth	T1	T2	T3	T4
ExercisePerMonth	T1	T2	T3	
PainNonActiveDaysPerMonth	T2	T3	T4	
WorkHoursPerWeek	T1	T2		
DrinkPerMonth	T1	T4		
HoursSleepPerDay	T2	T3		
AgeOfDiabetes	T1			
AnxiousDaysPerMonth	T3			
PhonesUsage	T4			
PrediabetesBloodSugar	T4			

Fig. 2: **Top five selected obesity risk factors for each subpopulation using MTL based on various ages, i.e., T1 (Young adults: $18 \leq age \leq 44$), T2 (Middle-aged adults: $45 \leq age \leq 64$), T3 (Older-aged adults: $65 \leq age \leq 99$) and T4 (Age information is missing). Note that, first column with variable names presents the names of obesity risk factors.**

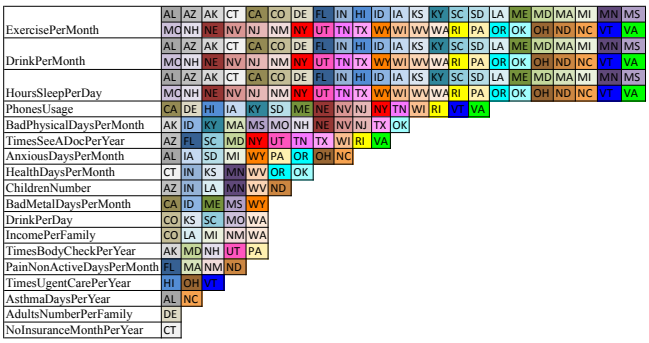

Fig. 3: **Top five selected obesity risk factors for each subpopulation using MTL based on geographic information, i.e., 46 states of the U.S. in the selected sample data set. States are represented by their corresponding abbreviations and each state is associated with one color.**

In the two MTL settings, the data set is divided into different number of subpopulations in terms of either age or geographic information. We then train a model for each subpopulation and these multiple models are trained simultaneously, so that we can obtain the ranked list of obesity risk factors for each subpopulation synchronously. In Fig. 2 and Fig. 3, subpopulation-level obesity risk factors can be located by linking the names of risk factors at the first column and each task with identical color. Besides the subpopulation-level obesity risk factors, we can also conclude the population-level obesity risk factors based on the shared obesity risk factors by all subpopulations. For example, the top three obesity risk factors in MTL2, shown in Fig. 3, are shared by all subpopulations, and thus are

TABLE I: **Top five selected obesity risk factors for all $2,000$ participants from two MTL settings (i.e., MTL1 and MTL2 are chosen from top five population-level obesity risk factors shown in Fig. 2 and Fig. 3, respectively.) and two STL methods (Linear model using generalized least squares (LMGLS) and linear mixed-effects model (LMEM)). Note that, the first column is the ranking number of the obesity risk factors.**

Ranking	MTL1	MTL2	LMGLS	LMEM
1	EnergyDaysPerMonth	ExercisePerMonth	DrinkPerDay	BadPhysicalDaysPerMonth
2	ExercisePerMonth	DrinkPerMonth	Age	BetterHealth
3	PainNonActiveDaysPerMonth	HoursSleepPerDay	AgeOfDiabetes	HadStroke
4	WorkHoursPerWeek	PhonesUsage	IncomePerFamily	ExercisePerMonth
5	DrinkPerMonth	BadPhysicalDaysPerMonth	ChildrenNumber	LimitedActivity

classified as population-level risk factors for obesity. This result also confirms the hypothesis that multiple tasks are related through sharing a common set of risk factors among the tasks, which is mentioned in Section I.

Because STL only can rank the obesity risk factors at the population-level, we compare the ranked list of obesity risk factors obtained and concluded from two MTL settings with the other the STL methods' lists in Table I. We select top five population-level obesity risk factors obtained from each method. Even we use two different ways of defining the tasks for MTL, the two ranked lists of obesity risk factors at the population-level share two obesity risk factors, i.e., ExercisePerMonth and DrinkPerMonth. However, the two selected STL methods' ranked list of risk factors do not share any obesity risk factor. As a result, we can see that MTL outperforms STL in terms of the capability of obtaining both subpopulation and population levels' ranked lists of obesity risk factors simultaneously as shown in Fig. 2 and Fig. 3.

IV. Conclusion

Risk factor analysis is a commonly used statistical method for the prevention, intervention and treatment of preventable diseases. On one hand, STL methods, including conventional regressions and standard statistical tests, learn the obesity risk factors ranking task at the population-level. On the other hand, MTL is a machine learning paradigm that leverages relatedness among the tasks to obtain the ranked list of obesity risk factors for each subpopulation, by simultaneously learning all the related tasks and transferring knowledge between tasks using $l_{2,1}-$norm regularization across all tasks. In our experiments, we compare two STL methods with MTL method, and MTL outperforms STL demonstrated in our experimental results that MTL is capable to rank the obesity risk factors at both the subpopulation and population-levels synchronously.

Albeit the MTL framework was presented to the subpopulation-level risk factors analysis, it is sufficiently flexible to be extended to solve the individual-level risk factor analysis. And hence, we will extend our present work by combining clustering technique to study the multilevel risk factor analysis simultaneously, i.e., individual and subpopulation levels.

Acknowledgment

This paper is based upon work supported by the National Science Foundation under grants CNS-1637312 and CCF-1451316.

References

[1] S. M. Grundy, "Obesity, metabolic syndrome, and cardiovascular disease," *The Journal of Clinical Endocrinology & Metabolism*, vol. 89, no. 6, pp. 2595–2600, 2004.

[2] A. Alwan *et al.*, *Global status report on noncommunicable diseases 2010.* World Health Organization, 2011.

[3] A. Chang, L. Van Horn, D. R. Jacobs, K. Liu, P. Muntner, B. Newsome, D. A. Shoham, R. Durazo-Arvizu, K. Bibbins-Domingo, J. Reis *et al.*, "Lifestyle-related factors, obesity, and incident microalbuminuria: the cardia (coronary artery risk development in young adults) study," *American Journal of Kidney Diseases*, vol. 62, no. 2, pp. 267–275, 2013.

[4] C. Attipa, K. Papasouliotis, L. Solano-Gallego, G. Baneth, Y. Nachum-Biala, E. Sarvani, T. G. Knowles, S. Mengi, D. Morris, C. Helps *et al.*, "Prevalence study and risk factor analysis of selected bacterial, protozoal and viral, including vector-borne, pathogens in cats from cyprus," *Parasites & vectors*, vol. 10, no. 1, p. 130, 2017.

[5] Y. Yang, X. Zhao, T. Dong, Z. Yang, Q. Zhang, and Y. Zhang, "Risk factors for postoperative delirium following hip fracture repair in elderly patients: a systematic review and meta-analysis," *Aging clinical and experimental research*, vol. 29, no. 2, pp. 115–126, 2017.

[6] J. E. Gangwisch, D. Malaspina, B. Boden-Albala, and S. B. Heymsfield, "Inadequate sleep as a risk factor for obesity: analyses of the nhanes i," *Sleep*, vol. 28, no. 10, pp. 1289–1296, 2005.

[7] M. A. Martínez-González, A. García-Arellano, E. Toledo, J. Salas-Salvado, P. Buil-Cosiales, D. Corella, M. I. Covas, H. Schröder, F. Arós, E. Gómez-Gracia *et al.*, "A 14-item mediterranean diet assessment tool and obesity indexes among high-risk subjects: the predimed trial," *PloS one*, vol. 7, no. 8, p. e43134, 2012.

[8] D. L. Smith, P. C. Fehling, A. Frisch, J. M. Haller, M. Winke, and M. W. Dailey, "The prevalence of cardiovascular disease risk factors and obesity in firefighters," *Journal of obesity*, vol. 2012, 2012.

[9] C. L. Ogden, M. D. Carroll, B. K. Kit, and K. M. Flegal, "Prevalence of childhood and adult obesity in the united states, 2011-2012," *Jama*, vol. 311, no. 8, pp. 806–814, 2014.

[10] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Advances in neural information processing systems*, 2007, pp. 41–48.

[11] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint l2,1-norms minimization," in *Advances in neural information processing systems*, 2010, pp. 1813–1821.

[12] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient l2, 1-norm minimization," in *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press, 2009, pp. 339–348.

[13] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[14] J. Zhou, J. Chen, and J. Ye, "Malsar: Multi-task learning via structural regularization," *Arizona State University*, vol. 21, 2011.

[15] J. Pinheiro, D. Bates, S. DebRoy, and D. Sarkar, "R core team (2014) nlme: linear and nonlinear mixed effects models. r package version 3.1-117," *Available at h ttp://CRAN. R-project. org/package= nlme*, 2014.

[16] U. N. S. Division, *Designing household survey samples: practical guidelines*. United Nations Publications, 2008, vol. 98.

[17] L. M. Howden and J. A. Meyer, "Age and sex composition: 2010," *2010 Census Briefs, US Department of Commerce, Economics and Statistics Administration. US CENSUS BUREAU*, 2010.