The Thirty-Third International
FLAIRS Conference (FLAIRS-33)

Not All Samples Are Equal: Class Dependent Hierarchical Multi-Task Learning for Patient Diagnosis Classification

Salim Malakouti,[1] **Milos Hauskrecht**[2]
Computer Science, University of Pittsburgh,
210 S Bouquet St.
Pittsburgh, PA 15213
salimm@cs.pitt.edu[1], milos@cs.pitt.edu[2]

Abstract

An interesting machine learning problem is to learn predictive models that can automatically assign diagnoses or diagnostic categories to patient cases. However, we often do not have enough positive samples for many of diagnoses either due to their rare nature or a limited size of available datasets. This motivates the use of multi-task learning methods that tend to improve model performance by imposing model similarities between related tasks. In this work, we tackle this important problem by exploring the benefits of existing expert-defined diagnostic hierarchies. We argue that related tasks (models) organized in expert-defined hierarchies do not have the same level of similarity for different classes of samples. We discuss how task similarities will be different for positive and negative samples and between parent and child diagnoses. We propose a new asymmetric version of Hierarchical Adaptive Multi-task Learning (HA-MTL) method that allows models to learn separate relatedness coefficient for tasks in the hierarchy based on their class values. Finally, we show that our model outperforms individually trained SVM models and symmetric HA-MTL results.

Introduction

Large scale adoption of Electronic Health Records (EHR) has facilitated many interesting problems such as automatic classification of patient diagnoses and diagnostic categories. Automatic classification and assignment of diagnoses at different levels of abstraction during patient's hospitalization is extremely useful for concisely summarizing the patient's condition. Additionally, it helps one explaining the patient's condition, as well as defining a proper context for choosing future patient management actions, or for supporting prediction and estimation of future outcomes.

However, a problem one often encounters when learning such classification models from data is that the number of positive samples available for some of the diagnoses (diagnostic codes) is very small. This is either due to the rare nature of the diseases or a limited size of available datasets. This motivates the use of multi-task learning methods that can take advantage of available data for similar diagnostic tasks. This is achieved by imposing similarities between behaviour of related models. However, past work has shown that multi-task learning methods can result in negative transfer if task similarities are not properly formulated (Rosenstein et al. 2005; Pan and Yang 2010) Therefore, past research work have proposed methods that model relationships between tasks by learning tasks relatedness coefficients (Ben-David and Schuller 2003; Kang, Grauman, and Sha 2011), creating task clusters, or by considering hierarchical structure of tasks (Jacob, Vert, and Bach 2009; Zhou, Chen, and Ye 2011a; Kumar and Daume III 2012).

In this paper we argue that it is important not only to consider which tasks to transfer from but also when to transfer from them as similarity does not imply that models' behaviour will be similarly related for all examples. In hierarchical settings such as patient diagnoses classification problem, parent diagnostic categories include similar groups of diagnostic codes. Therefore, there will be similarities between the diagnostic parent's predictions and its children. This has motivated methods that allow transfer of model parameters in a top-down fashion (Dumais and Chen 2000; Wu, Zhang, and Honavar 2005). However, one can argue that this adaptation may not be symmetric for positive and negative classes (in binary classification tasks). In other words, it is intuitive to think that negative class label of the parent diagnostic task is more likely to translate to a negative class label of its particular children, however, it is not possible to make the same claim for positive instances. In fact we often expect positive class label of the parent to imply that at least one but not all child tasks to have positive class as well. Therefore, imposing same weight for the parent diagnostic model for both classes during a top-down learning mechanism seems counter intuitive. One would expect that the relatedness of parent and child tasks to be asymmetric instead, but to the extent of our knowledge none of the existing work have studied this in hierarchical multi-task learning settings.

In the rest of this paper, we first provide introductory information on Hierarchical Adaptive Multi-task Learning (HA-MTL) method (Malakouti and Hauskrecht 2019b). Next we study how HA-MTL is imposing similarities among the tasks in top-down and bottom-up transfer of model parameters. Next, we propose a new hierarchical version of HA-MTL method that allows asymmetric class dependent

adaptation of model behaviours by learning class specific relatedness coefficients. Finally, we show our method can learn models with improved classification performance and analyze the difference between model adaptation from parent diagnostic categories for positive and negative classes.

Related Work

The problem of modelling patient wide range of diagnoses has recently become a focus of machine learning research. Some of existing work tackle the problem of predicting future admission diagnoses using past history of patients (Lipton et al. 2015; Choi et al. 2017) while others have taken on the challenge of assigning diagnostic codes to current patient's visit (Pakhomov, Buntrock, and Chute 2006; Miotto et al. 2016). Recently new methods have been proposed to take advantage of hierarchical structure of patient diagnoses to expand the training features (Choi et al. 2017). Other methods have attempted to leverage task hierarchies to improve learning of diagnostic models using hierarchical classification methods (Perotte et al. 2013) and finally new methods have been proposed to use the hierarchical relationships of diagnostic codes in multi-task learning settings (Malakouti and Hauskrecht 2019a).

Hierarchical Multi-task learning for large number of tasks is closely related to the two types of machine learning methods: Hierarchical Learning and Multi-task Learning itself. In the hierarchical learning community top-down learning of machine learning models have been commonly studied before (Koller and Sahami 1997). In the top-down approach a classifier on a low-level of the hierarchy is defined using a decision or a signal generated by its parent classifiers. There are different versions of the top-down approach that place various consistency constraints on predictions of the parent and child tasks and their classifier outputs, most frequently assuring the probability of a parent (diagnostic category) is higher than the probability of a low-level class or class category (Dumais and Chen 2000; Wu, Zhang, and Honavar 2005). The main problem with the top-down approach is that learning of higher level class models from data may omit details that the low-level class models could capture. For example, some of the findings for a patient may point specifically and with a high accuracy to a low-level diagnosis while the higher level class model marginalizes it out during the learning and as a result does not include it in the model. In such a case the probability of a lower-level class may be higher than the probability of a higher level class category violating the constraint consistency. One way to correct for child-to-parent effects is to define and add a bottom-up process that assures positive lower-level class predictions aggregate properly in the parent tasks (Valentini 2011). However, pure bottom-up approach would require the presence of accurate classifier models on the leaf classification layer, which is hard to achieve in practice when datasets of a limited size are used to train such models and the count of positive instances for such classes are very low. There exists a variety of hierarchical classification methods that try to account for both the top-down and bottom-up classification processes. One example is a Bayesian aggregation method by (DeCoro and Barutcuoglu 2006) that compiles the hierarchy into the Bayesian belief network and uses inferences to support the classification on a different level of hierarchy. Other methods that account for both top-down (negative) and bottom-up (positive) effects are based on structured margin classifiers (Dekel, Keshet, and Singer 2004).

Multi-task learning methods tackle this problem by devising machine learning algorithms that tend to improve individual task performance by simultaneously learning of all tasks (Zhang and Yang 2017). These methods tend to promote sharing of information between related tasks by imposing similarities between them. Evgeniou and Pontil proposed a multi-task learning method based on SVM algorithm that learns all tasks simultaneously by regularizing their differences from their average (Evgeniou and Pontil 2004). Other methods have been proposed to tackle this problem by performing shared feature learning (Argyriou, Evgeniou, and Pontil 2007; 2008).

It has been shown that multi-task learning methods are highly sensitive to relatedness of tasks, and if not modeled properly, they can result in negative transfer (Rosenstein et al. 2005; Pan and Yang 2010). One approach to solve this problem is to explicitly model and learn task relationships (Ben-David and Schuller 2003; Kang, Grauman, and Sha 2011; Zhang and Yeung 2012; Zhou, Chen, and Ye 2011b). The second group of methods attempted to learn task groups or clusters to prevent negative transfer from unrelated tasks (Jacob, Vert, and Bach 2009; Zhou, Chen, and Ye 2011a; Kumar and Daume III 2012). However, none of these methods take advantage of the complex hierarchical relationships between tasks when they are organized in hierarchies. Some earlier work have studied and proposed hierarchical multi-task learning problem in which task relationships cannot simply be modeled using flat task clusters (Liu et al. 2017; Kim and Xing 2010; Xue et al. 2007; Malakouti and Hauskrecht 2019a). However, there are yet many open questions to be answered. One interesting question is to investigate whether positive transfer of model parameters between tasks happen equally across different classes. In other words, are the negative samples and positive samples equally important in imposing similarities between related tasks.

In the rest of this section we first review HA-MTL method and propose an asymmetric class dependent extension for it called asymmetric HA-ML. Next we evaluate our method's performance on the patient diagnosis classification task and study class-dependent transfer weights.

Methodology

Assume we have T diagnoses and diagnostic categories, each covered by a separate binary classification task. The tasks are organized in a hierarchical structure H. Additionally, we assume each patient's $X \in R^D$ consists a D dimensional dense representation of patient's EHR data. Our objective is to learn T discriminant functions $f_1, f_2, ..., f_T$ in which $f_t : \mathbb{R}^D \rightarrow \mathbb{R}$. Hence, the predicted score of the discriminant function f_t can be mapped to one of the binary labels $0, 1$ using a task specific threshold γ_t. Finally, we introduce a $\phi t, H$ as the parent of tasks t in the hierarchy H.

The conventional method is to learn the T discriminant functions independently. However, recent work has shown that by adapting model parameters in a top-down fashion one can improve lower level diagnostic models by taking advantage of more general diagnostic models trained for their parent diagnostic categories. In this section, we first formalize Hierarchical Adaptive Multi-task Learning method (Malakouti and Hauskrecht 2019a). Next, we show that the transfer of model parameters is achieved by minimizing the difference between the target task parameters and a weighted sum of auxiliary tasks. Finally, we propose a new class dependent extension of HA-MTL that allows class dependent asymmetric adaptation from parent diagnostic models.

HA-MTL: Hierarchical Adaptive Multi-task Learning

HA-MTL's goal is to adapt model parameters from parent and child diagnostic tasks while simultaneously learning the importance of the set of auxiliary models. They define the set of auxiliary tasks for the target task t as the set of its parent and child diagnostic tasks. HA-MTL improves each diagnostic task in an iterative fashion by proposing a two phase (top-down and bottom-up) adaptation algorithm that transfers model parameters from either their parent or child diagnostic models. In order to perform the model parameter adaptation and simultaneously learn the importance of auxiliary task they propose Regularized Adaptive SVM (RA-SVM) as show in equation 1.

$$\begin{aligned} \min_{v_t,\varepsilon,\tau} \quad & \sum_{i}^{N_t} \varepsilon_i + C_1||v_t||^2 + C_2||\tau||^2 \\ s.t. \quad & y_i \sum_{a \in axu(t)} \tau_a f_a(x_i) + y_i v_t^T x_i \geq 1 - \varepsilon_i \\ & i = 1, ..., N_t, \quad \varepsilon_i \geq 0 \end{aligned} \tag{1}$$

Where v_t corresponds to the model parameters for $\Delta f_t = f_t - \sum_a^{aux(t)} \tau_a f_a$ and τ_a refers to the relatedness or usefulness coefficient of auxiliary task a for target task t. The function $aux(t)$ provides the set of auxiliary tasks of t in the corresponding top-down or bottom-up step. Finally, values of C_1 and C_2 determine the trade-off between regularizing model parameters and auxiliary task weights. The optimization problem in Equation 1 is minimizing the hinge loss while also regularizing both the auxiliary task weights and model parameters of Δf_t.

RA-SVM improves upon Adaptive SVM algorithm first proposed by Yang et al.(Yang, Yan, and Hauptmann 2007). The advantage of RA-SVM is two fold. First, it attempts to simultaneously learn the importance of each auxiliary task, while, the original A-SVM method relied on input to provide this information. Second, RA-SVM can be optimized using any standard SVM library by relaxing the assumption that $\sum_{a \in aux(t)} \tau_a = 1$ in which τ_a is the influence of auxiliary task a. This is done be defining a feature map over the original features and prediction scores of auxiliary tasks.

Lemma 1. *RA-SVM finds a trade off between imposing similarities between the predictions of the target task and the weighted average predictions of the auxiliary task models while regularizing auxiliary task weights or in other words learning the target task independently.*

This can be shown by re-writing the optimization problem in Equation 1 and by replacing $f_a(x_i)$ with ${w_a}^T x_i$ assuming all auxiliary tasks are trained using linear models. Additionally, the final models parameters for task t can be written as $w_t = \sum_a \tau_a {w_a}^T x_i + v_t$. Therefore, Equation 1 can be re-written as:

$$\begin{aligned} \min_{v_t,\varepsilon,\tau} \quad & \sum_{i}^{N_t} \varepsilon_i + C_1||w_t - \sum_{a \in axu(t)} \tau_a {w_a}^T||^2 \\ & + C_2||\tau||^2 \\ s.t. \quad & y_i w_t^T x_i \geq 1 - \varepsilon_i \\ & i = 1, ..., N_t, \quad \varepsilon_i \geq 0 \end{aligned} \tag{2}$$

The term $||w_t - \sum_a \tau_a {w_a}^T||$ in Equation 2 attempts to regularize the difference between target task model outcomes from the weighted average of the auxiliary tasks. This is while $||\tau||^2$ promotes smaller influence of auxiliary tasks. This creates a trade-off between masking the impact of unnecessary auxiliary tasksand imposing similarities between the target task and chosen auxiliary tasks. In fact high values of $\frac{C1}{C2}$ promotes further regularization of $||w_t - \sum_a \tau_a {w_a}^T||$ and therefore promotes higher impact of auxiliary weights. On the other hand lower ratios of $C1$ and $C2$ promote independent learning of target task.

Not All Samples are Equal

RA-SVM method assumes the signal from auxiliary tasks are equally useful in improving target task model's performance. However, intuitively one can imagine that auxiliary model scores in a hierarchical structure can have different meaning or impact based on the type of the dependency between related tasks. For instance, in the top-down adaptation phase a negative score of parent model (assuming parent model has a higher performance) is more likely to translate to a negative label for the child task. In contrast, a positive class prediction of the parent may not necessarily mean that the child task will also be positive. As previously discussed in the hierarchical classification literature (Silla and Freitas 2011) negative samples in a hierarchy are passed top-down while positive class samples are promoted in a bottom-up fashion. Therefore, in this work, we propose an asymmetric adaptation mechanism based on ReLU operation to break down the scores of RASVM models to a pair of class dependent signals $f_a^p = max(0, f_a)$ and $f_a^p = min(0, f_a)$.

The signals $f_a^p \in [0\ \infty]$ and $f_a^n \in [-\infty\ 0]$ allow target task models to learn two different relatedness coefficients τ_a^p and τ_a^n for positive and negative signals from auxiliary tasks. RA-SVM optimization problem in Equation 1 can be written for AsymRA-SVM as shown in equation below:

$$\min_{v_t,\varepsilon,\tau} \quad \sum_{i}^{N_t} \varepsilon_i + C_1||v_t||^2 + C_2||\tau||^2$$
$$s.t. \quad y_i \sum_{a \in axu(t)} \tau_a^p f_a^p(x_i) + \tau_a^n f_a^n(x_i) \quad (3)$$
$$+ y_i v_t^T x_i \geq 1 -$$
$$\varepsilon_i i = 1, ..., N_t, \quad \varepsilon_i \geq 0$$

AsymRA-SVM enables the target task model to learn how important each signal from auxiliary tasks is by minimizing $||w_t - \sum_a \tau_a^p f_a^p - \tau_a^n f_a^n||$ and learning two separate weights for each auxiliary task.

Optimizing Prediction Thresholds

AsymRA-SVM splits the auxiliary task signals to a positive and negative signal. However, this can often be an issue since the optimum decision threshold γ_t may not always be zero. In order to address this problem we attempt to optimize the decision threshold γ_t by maximizing the F1 score. Since F1 is neither differential nor a convex function (Busa-Fekete et al. 2015) we used Particle Swarm Optimization method which has been shown to be suitable for ill-formatted, non-differentiable and non-convex optimization problems (Shi and Eberhart 1999; Park et al. 2009). To allow f_a^p and f_a^n to remain in $[0\ \infty]$ and $[-\infty\ 0]$ ranges we redefined f_t as $f_t = w_t + b_t + \gamma_t$.

Experiments

We conducted our experiments on 22046 ICU patient admissions from MetaVision subset of MIMIC-III dataset (Johnson et al. 2016). MIMIC III provides patient diagnosis using International Classification of Diseases (ICD-9)(Slee 1978) codes. Therefore, we rely on ICD-9 hierarchy to obtain diagnostic tasks and their hierarchical structure. We obtained categorical ground truth labels by applying a logical OR operation between all its children. Finally, we omitted any diagnostic code with less than 30 positive samples (imbalance ratio of 0.001) while keeping 2019 diagnostic codes as learning tasks.

The problem of learning a dense representation of patient EHR data as $X \in R^D$ has been studied in past (Miotto et al. 2016; Hauskrecht et al. 2016). In this paper we followed the method proposed in the original HA-MTL paper (Malakouti and Hauskrecht 2019a). We used Latent Semantic Indexing(LSI) method which uses singular value decomposition(SVD) to learn a lower dimensional representation of the admission-event matrix. In order to create admission-event matrix to be used by SVD, we converted patient's EHR data to a set of meaningful binary events by converting medication and procedure orders to indicators events. Laboratory results and physiological measurements with numerical values were converted to Normal and Abnormal Low or High events based on their standard normal ranges. Finally, Discrete valued measurements and pain level assessments were also converted to specific events matching each unique value. After the conversion, our new EHR events data consisted of 4826 clinical events. This includes 2420 for medication orders, 116 for procedure orders, 2012 for laboratory results and 278 for physiological and pain assessment measurements. Finally, we created admission-event matrix by obtaining a BoW representation of patient events data with normalized frequencies.

Table 1: Average performance for all diagnostic tasks

Method Name	**AUROC**	**F1**
Random	0.5	0.0247
SVM	0.764	0.0755
HA-MTL$_{td}$	0.770	0.110
AsymHA-MTL$_{td}$	**0.778**	**0.135**

We evaluate our method by comparing the performance of AsymmHA-MTL to three baselines including random guessing model, individually trained SVM models and the original HA-MTL algorithm. We use used Area Under Receiver Operating Curve (AUROC) and F1 score to provide quantitative comparison of the methods over a random stratified (70%-30%) split of data to train and test set. Finally, we used random sub-sampling to generate 10 different 75%/25% train/validation splits from the train set for hyper parameters optimization.

Table 1 shows the average AUROC and F1 for all 2019 diagnostic tasks. AsymmHA-MTL method is outperforming the baselines and symmetric HA-MTL method. However as discussed in (Malakouti and Hauskrecht 2019a) we expect the majority of improvements to happen on lower level child diagnosis. Therefore, in Table 2 we have provided results for parts of two sub branches of ICD-9 heirarchy for Fracture of ribs and Diabetes. Our results show that considerable improvements has been gained by both HA-MTL and AsymmHA-MTL while AsymmHA-MTL is outperforming others. Similar trends are visible across different sub branches of the hierarchy.

Table 3 shows the improvements (AUROC) of models across different ranges of task priors or $P(y_i = 1)$. AsymmHA-MTL is outperforming HA-MTL across different ranges of tasks priors while both methods have higher gains in tasks with fewer number of positive samples.

Despite the improvements of HA-MTL and AsymmHA-MTL, one negative observation is that both HA-MTL and AsymmHA-MTL is that in many case RASVM methods fail to prevent negative transfer from auxiliary tasks. This is mor common in tasks with very high imbalance ratio (near 0.001) and small number of positive samples. Analysis of tasks results suggests that this is because RA-SM sometimes fails to prevent negative transfer by failing to choose right values of $C1$ and $C2$ hyperparamters using internal cross-validation. This can be explained by the significantly small number of positive samples in validation and test set.

Figure 1 depicts the distribution of tau_p (impact of positive signals) and tau_n (impact of negative signals) in a top-down transfer of model parameters. The contrast between the distribution of positive and negative signals show that models are more likely to learn a higher impact for f_a^n when

Table 2: Comparison of methods for example branches of ICD9

Diagnostic Task Name	**SVM AUROC**	**HA-MTL AUROC**	**AsymmHA-MTL AUROC**
Diabetes mellitus	0.866	0.863	0.864
_ Diabetes mellitus without mention of complication	0.714	0.718	**0.773**
___ Diabetes mellitus without mention of complication, type II	0.689	0.68	**0.757**
_ Diabetes with hyperosmolarity	0.774	**0.863**	**0.858**
___ Diabetes with renal manifestations, type II	0.805	0.823	**0.852**
Fracture of rib(s) sternum larynx and trachea	0.914	0.912	0.919
_ Closed fracture of rib(s)	0.90	0.911	0.915
___ Closed fracture of multiple ribs, unspecified	0.690	0.764	**0.833**

Table 3: Average model improvements (AUROC) for diagnostic tasks within different ranges of prior for positive class

Average model improvements	**0 - 0.0016**	**0.0016 - 0.003**	**0.003 - 0.01**	**0.01 - 0.05**	**0.05 - 0.1**	**0.1 - 0.5**	**0.5 - 1**
Number of tasks	91	467	763	495	95	100	8
AsymmHA-MTL	**0.0252**	**0.0158**	**0.0092**	0.0051	0.0001	-0.0003	0.0002
HA-MTL	0.0180	0.0095	0.0055	0.0030	-0.0011	-0.0018	0.0001

a is a parent of target task. This agrees with our hypothesis discussed in section that negative signals from parent diagnostic tasks are more likely to translate to a negative score in the child diagnostic model.

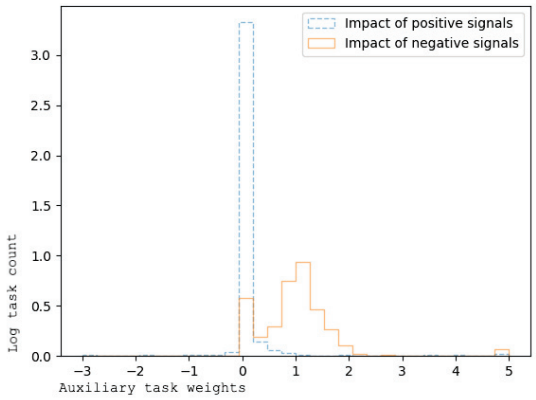

Figure 1: Distribution of τ_p and τ_a values in the top-down transfer of model parameters

Conclusion

In this paper we argued that usefulness and impact of related tasks in hierarchical multi-task learning problems can depend not only to the tasks but also on the classes of samples. For example in the top-down transfer of parameters high negative scores of parent models is more likely to translate to negative scores of lower child models as negative labels are passed from a parent to child while this is not necessary true for positive labels. Therefore, we proposed an asymmetric hierarchical adaptive multi-task learning method that allows models to simultaneously learn model parameters and importance of positive and negative scores of auxiliary tasks independently. Our results show that during a top-down model adaptation phase our model is able to improve model performances compared to symmetric version of the algorithm and baseline SVM models.

References

Argyriou, A.; Evgeniou, T.; and Pontil, M. 2007. Multi-task feature learning. In *Advances in neural information processing systems*, 41–48.

Argyriou, A.; Evgeniou, T.; and Pontil, M. 2008. Convex multi-task feature learning. *Machine Learning* 73(3):243–272.

Ben-David, S., and Schuller, R. 2003. Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines*. Springer. 567–580.

Busa-Fekete, R.; Szörényi, B.; Dembczynski, K.; and Hüllermeier, E. 2015. Online f-measure optimization. In *Advances in Neural Information Processing Systems*, 595–603.

Choi, E.; Bahadori, M. T.; Song, L.; Stewart, W. F.; and Sun, J. 2017. Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 787–795. ACM.

DeCoro, C., and Barutcuoglu, Z. 2006. Hierarchical shape classification using bayesian aggregation. In *IEEE International Conference on Shape Modeling and Applications 2006(SMI)*, volume 00, 44.

Dekel, O.; Keshet, J.; and Singer, Y. 2004. Large margin hierarchical classification. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, 27–. New York, NY, USA: ACM.

Dumais, S., and Chen, H. 2000. Hierarchical classification of web content. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Develop-*

ment in Information Retrieval, SIGIR '00, 256–263. New York, NY, USA: ACM.

Evgeniou, T., and Pontil, M. 2004. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 109–117. ACM.

Hauskrecht, M.; Batal, I.; Hong, C.; Nguyen, Q.; Cooper, G. F.; Visweswaran, S.; and Clermont, G. 2016. Outlier-based detection of unusual patient-management actions: an icu study. *Journal of biomedical informatics* 64.

Jacob, L.; Vert, J.-p.; and Bach, F. R. 2009. Clustered multi-task learning: A convex formulation. In *Advances in neural information processing systems*, 745–752.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Li-wei, H. L.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data* 3:160035.

Kang, Z.; Grauman, K.; and Sha, F. 2011. Learning with whom to share in multi-task feature learning. In *ICML*, volume 2, 4.

Kim, S., and Xing, E. P. 2010. Tree-guided group lasso for multi-task regression with structured sparsity. In *ICML*, volume 2, 1.

Koller, D., and Sahami, M. 1997. Hierarchically classifying documents using very few words. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, 170–178. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Kumar, A., and Daume III, H. 2012. Learning task grouping and overlap in multi-task learning. *arXiv preprint arXiv:1206.6417*.

Lipton, Z. C.; Kale, D. C.; Elkan, C.; and Wetzel, R. 2015. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint*.

Liu, A.-A.; Su, Y.-T.; Nie, W.-Z.; and Kankanhalli, M. 2017. Hierarchical clustering multi-task learning for joint human action grouping and recognition. *IEEE transactions on pattern analysis and machine intelligence* 39(1):102–114.

Malakouti, S., and Hauskrecht, M. 2019a. Hierarchical adaptive multi-task learning framework for patient diagnoses and diagnostic category classification. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE.

Malakouti, S., and Hauskrecht, M. 2019b. Predicting patient's diagnoses and diagnostic categories from clinical-events in ehr data. In *Conference on Artificial Intelligence in Medicine in Europe*, 125–130. Springer.

Miotto, R.; Li, L.; Kidd, B. A.; and Dudley, J. T. 2016. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports* 6:26094.

Pakhomov, S. V.; Buntrock, J. D.; and Chute, C. G. 2006. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association* 13(5):516–525.

Pan, S. J., and Yang, Q. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22(10):1345–1359.

Park, J.-B.; Jeong, Y.-W.; Shin, J.-R.; and Lee, K. Y. 2009. An improved particle swarm optimization for nonconvex economic dispatch problems. *IEEE Transactions on Power Systems* 25(1):156–166.

Perotte, A.; Pivovarov, R.; Natarajan, K.; Weiskopf, N.; Wood, F.; and Elhadad, N. 2013. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association* 21(2):231–237.

Rosenstein, M. T.; Marx, Z.; Kaelbling, L. P.; and Dietterich, T. G. 2005. To transfer or not to transfer. In *NIPS 2005 workshop on transfer learning*, volume 898, 1–4.

Shi, Y., and Eberhart, R. C. 1999. Empirical study of particle swarm optimization. In *Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406)*, volume 3, 1945–1950. IEEE.

Silla, C., and Freitas, A. 2011. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery* 22:31–72.

Slee, V. N. 1978. The international classification of diseases: ninth revision (icd-9). *Annals of internal medicine* 88(3):424–426.

Valentini, G. 2011. True path rule hierarchical ensembles for genome-wide gene function prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 8(3):832–847.

Wu, F.; Zhang, J.; and Honavar, V. 2005. Learning classifiers using hierarchically structured class taxonomies. In *Lecture Notes in Computer Science*, volume 3607. Springer, Germany.

Xue, Y.; Liao, X.; Carin, L.; and Krishnapuram, B. 2007. Multi-task learning for classification with dirichlet process priors. *Journal of Machine Learning Research* 8(Jan):35–63.

Yang, J.; Yan, R.; and Hauptmann, A. G. 2007. Cross-domain video concept detection using adaptive svms. In *Proceedings of the 15th ACM international conference on Multimedia*, 188–197. ACM.

Zhang, Y., and Yang, Q. 2017. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*.

Zhang, Y., and Yeung, D.-Y. 2012. A convex formulation for learning task relationships in multi-task learning. *arXiv preprint arXiv:1203.3536*.

Zhou, J.; Chen, J.; and Ye, J. 2011a. Clustered multi-task learning via alternating structure optimization. In *Advances in neural information processing systems*, 702–710.

Zhou, J.; Chen, J.; and Ye, J. 2011b. Malsar: Multi-task learning via structural regularization. *Arizona State University* 21.