Literature Mining for Incorporating Inductive Bias in Biomedical Prediction Tasks (Student Abstract)

Qizhen Zhang,[1,3] **Audrey Durand,**[2,3] **Joelle Pineau**[1,3]
[1]McGill University, [2]Université Laval, [3]Mila – Quebec AI Institute
3480 University St., Montreal, Quebec H3A 0E9
Phone: (514)-398-5432
E-mail: qizhen.zhang@mail.mcgill.ca

Abstract

Applications of machine learning in biomedical prediction tasks are often limited by datasets that are unrepresentative of the sampling population. In these situations, we can no longer rely only on the the training data to learn the relations between features and the prediction outcome. Our method proposes to learn an inductive bias that indicates the relevance of each feature to outcomes through literature mining in PubMed, a centralized source of biomedical documents. The inductive bias acts as a source of prior knowledge from experts, which we leverage by imposing an extra penalty for model weights that differ from this inductive bias. We empirically evaluate our method on a medical prediction task and highlight the importance of incorporating expert knowledge that can capture relations not present in the training data.

Introduction

Supervised learning models have become a popular tool for solving biomedical prediction problems (Coudray et al. 2018, e.g.) due to their good performance and to the availability of new datasets. While training these models, regularization (e.g. L1 (Tibshirani 1996) and L2 (Hoerl and Kennard 1970)) is commonly used to prevent model overfitting. However, these approaches are confined in the sense that *their knowledge is limited to what can be extracted from the given data*. This can be a problem when the dataset is not representative of the sampling population. The "healthy volunteer effect" (Fry et al. 2017) is a common example, where people volunteering for health-related research studies tend to be more health-conscious than nonparticipants. The information contained in such dataset would therefore be incomplete for training a model to predict in the population of interest. In addition, resulting models could be hard to interpret if they failed to capture dynamics expected by experts. This motivates us to leverage prior knowledge about feature and outcome relations, which has already been well studied by experts in the biomedical disciplines.

The general idea is to take advantage of prior knowledge extracted from literature mining in PubMed[1], a search engine containing 29 million biomedical documents from various sources including life science journals. We suggest to learn how important each feature is for the prediction task by querying PubMed and use this as an inductive bias on the learned model. We investigate ways to incorporate such an inductive bias (corresponding to expert knowledge) in the regularization process. Regularization with prior knowledge is not a new idea, for example, L1 and L2 can be interpreted as having a Laplace and Gaussian prior respectively (Robert 2014). In this work, we propose an approach which leverage the expert knowledge using a regularization term aiming to minimize the Kullback-Leibler (KL) divergence between the model weights distribution and the inductive bias distribution. We evaluate this approach on a medical prediction task to highlight the importance of incorporating expert knowledge that can capture relations not present in the training data.

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]https://www.ncbi.nlm.nih.gov/pubmed/

Literature Mining of Inductive Bias

PubMed was designed to be queried as a centralized knowledge source for biomedical experts, thus appearing as an appropriate source of expert knowledge for biomedical prediction tasks. We obtain our expert knowledge in two steps: 1) extracting relevant papers and 2) quantifying the relation between each feature and the outcome.

Step 1) We first find the corresponding Medical Subject Headings (MeSH) in PubMED for the task. A MeSH would contain documents relevant to the prediction task, and can be easily obtained through PubMed's API.

Step 2) The relation between a feature and the outcome is quantified by counting occurrences of its keywords in all papers. More specifically, we first obtain a set $\mathcal{S}_i$ which contains k_i keywords [2] for feature i using medical concept extraction tool quickUMLS (Soldaini and Goharian 2016). Then, for each feature i, we sum the number of occurrences of each keywords in $\mathcal{S}_i$ retrieved in all the abstracts and titles that subside in the selected MeSH. These number of occurrences, which we denote α_i, serve as a proxy for measuring the relevancy between the feature and the prediction task. At this point, we have a vector $\boldsymbol{\alpha} = (\alpha_i)_{i=1,..N}$ that contains

[2]See Supp. for details and benefits of this approach.

a *relevancy score* for each of the features, where N denotes the number of features in the dataset. Now, we want to incorporate that knowledge within a classifier.

Regularizing with Inductive Bias

We aim at preventing the learned model from diverging too much from the expert knowledge. To this end, we use a KL divergence penalization on the *distribution of attention over features* learned by the model and the one obtained through literature mining. These comparable distributions are obtained by transforming the model weights w_i and relevance scores, respectively.

Transforming model weights The normalized model weights distribution $\widetilde{\mathbf{Z}} = [\widetilde{z_1}, \ldots, \widetilde{z_N}]$ over features is computed as follows. We take into account cases where categorical feature are one-hot encoded by taking the average:

$$\widetilde{z_i} = \frac{r_i}{\sum(r_i)}, \text{where} \tag{1}$$

$$r_i = \begin{cases} |w_i| & \text{if feature } i \text{ was not one hot encoded;} \\ \frac{\sum_{j \in o_i} |w_j|}{|o_i|} & \text{if } o_i \text{ is the set of one hot encodings for feature } i. \end{cases}$$

Note that we assume a model which learns exactly one scalar weight for each encoded feature. However, the above can also be easily extended to models learning $M \geq 1$ weights for each encoded feature by taking the average in Eq. 1.

Transforming relevancy scores We compute a normalized relevancy distribution $\mathbf{Z} = [z_1, ..., z_N]$ over features as:

$$z_i = \frac{\alpha_i / k_i}{\sum_{j=1}^{N} (\alpha_j / k_j)}. \tag{2}$$

Inductive Bias Regularization Let $\mathcal{L}_E$ denote the error loss (e.g. binary cross-entropy loss) and let $\lambda \geq 0$ be a hyper-parameter, we propose the following regularized loss:

$$\mathcal{L} = \mathcal{L}_E + \lambda D_{KL}(\mathbf{Z} || \widetilde{\mathbf{Z}}). \tag{3}$$

Experiments

We investigate the effectiveness of the proposed strategy using experiments on a cardiovascular disease prediction dataset (Ulianova 2019) containing 70,000 individuals and 11 features[3]. We craft synthetic experiments where training data is such that the relationship between the most important feature (i.e. systolic blood pressure, SBP) and the outcome cannot be detected[4]. We then compare logistic regression models[5] with no regularization, traditional L1 regularization, and inductive bias regularized loss (Eq. 3). Training, validation[6], and testing sets contain 9237, 9237, and 51526 samples, respectively.

[3] See Supp. for datasets details.

[4] See Supp. for data split and SBP identification details

[5] When sufficiently accurate in terms of prediction, its coefficients are generally treated as a gold standard for interpretability.

[6] See Supp. for hyper-parameters details.

Table 1: Test accuracy on cardiovascular dataset

Model	Accuracy	learnt SBP weight
No regularization	0.323	-0.011
L1	0.323	-0.011
With inductive bias	**0.748**	**0.127**

Results

Table 1 shows the preliminary results. As expected, both no regularization and standard L1 regularization fail to capture the link (invisible from training data) between SBP and the outcome. With inductive bias, we are able to learn the relationship between the feature and outcome, resulting in significantly improved performance. Additional results investigating the benefits of using QuickUMLS instead of feature names directly are provided in Supp.

Conclusion and Future Works

We have proposed a new method for incorporating inductive bias obtained from literature mining. Preliminary empirical results show the potential of our approach when the dataset is unrepresentative of the sampled distribution. One limitation of the method is that it requires good descriptive identifiers for all features, which are then used to find their keywords. We also assume that each prediction task always has a corresponding MeSH in PubMed. These are not the case for all datasets and tasks.

As part of future works, we will apply the method to larger datasets, and investigate other ways to incorporate the inductive bias such as using the Wasserstein metric.

References

Coudray, N.; Ocampo, P. S.; Sakellaropoulos, T.; Narula, N.; Snuderl, M.; Fenyö, D.; Moreira, A. L.; Razavian, N.; and Tsirigos, A. 2018. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine* 24(10):1559.

Fry, A.; Littlejohns, T. J.; Sudlow, C.; Doherty, N.; Adamska, L.; Sprosen, T.; Collins, R.; and Allen, N. E. 2017. Comparison of sociodemographic and health-related characteristics of uk biobank participants with those of the general population. *American journal of epidemiology* 186(9):1026–1034.

Hoerl, A. E., and Kennard, R. W. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12(1):55–67.

Robert, C. 2014. Machine learning, a probabilistic perspective.

Soldaini, L., and Goharian, N. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, SIGIR*.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1):267–288.

Ulianova, S. 2019. Cardiovascular disease dataset.