High-Performing Machine Learning Models for Phenotype Development

Victor A. Rodriguez, MPhil[1*], Tony Y. Sun, BS[1*], Phyllis M. Thangaraj, MPhil[1*], Karthik Natarajan, PhD[1], Patrick Ryan, PhD[1] (* equal contribution)
Department of Biomedical Informatics, Columbia University, New York City, NY [1]

Abstract

Phenotyping algorithms are essential tools for conducting clinical research with observational data. Developing phenotypes in a high-throughput manner, however, remains difficult.[1] We address this challenge by learning how to build concept sets for novel phenotype definitions--a task which often demands time-consuming manual curation by expert clinicians and informaticians. Our approach leverages the structure of existent phenotype definitions to inform construction of novel phenotype concept sets. We use eMERGE validated PheKB phenotypes as a source of existent phenotype concept sets, and analyze their structure in terms of concept pairs.[3] Through our collaboration with the Phenotyping Working Group at Columbia's Department of Biomedical Informatics[2], we characterize concept pairs by their co-occurrence; their semantic, lexical, and embedding similarity; and a binary indicator encoding their presence or absence within any of our existent phenotype concept sets. We train various binary classifiers to predict if a given concept pair should appear within a phenotype concept set given its features. We hypothesize that these models can be utilized to rapidly construct novel phenotype concept sets in a scalable manner.

Methods

We work with clinical concepts from OMOP CDM standard vocabularies as defined by the Observational Health Data Sciences and Informatics (OHDSI) Program. Concept pairs are characterized using patient data from Columbia University Irving Medical Center's Clinical Data Warehouse (CUIMC CDW) organized in the OMOP CDM format.

We build our models with data for over 2.3×10^7 unique concept pairs, each described by 21 features belonging to one of four categories: lexical, semantic, co-occurrence, or co-occurrence-based embeddings[4]. We evaluate how well our models perform with respect to 1) random held-out concept pair prediction (random hold-out) and 2) prediction of concept pairs within fully held-out phenotype concept sets (phenotype-aware hold-out).

For random hold-out, we construct a held-out test set by randomly sampling 10% of our positive concept pairs as well as an equal number of negative concept pairs. The remaining 90% of positive pairs along with an equal number of randomly sampled negative pairs are used as the training set. We repeat this process ten times to generate ten random training and test sets.

For phenotype-aware hold-out, we randomly select ten existent phenotype concept sets whose positive concept pairs contain approximately 10% of all positive pairs. We hold out all positive pairs contained in these concept sets and randomly sample an equal number of negative pairs to create a test set. For the training set, we use all remaining positive pairs along with a matching number of randomly sampled negative pairs. As with random hold-out, we repeat this process ten times to generate ten phenotype-aware held-out training and test sets.

For each hold-out method, we train and test multiple binary classifiers and aggregate results over all our test sets. To determine the feature-weightings that influenced the performance of the models, we re-ran L1 logistic regression with 15 combinations of the 4 different groups of features, including every combination within each group (e.g. running permutations of semantic features, lexical features, etc) .

Results for concept pair prediction

Train/Test Split	Model	AUROC	AUPRC	Max. F1	Prec.@50%
Random	LR (L1)	0.9417 ± 0.0008	0.9337 ± 0.0014	0.8701 ± 0.0011	0.9700 ± 0.0205
	LR (L2)	0.9402 ± 0.0009	0.9321 ± 0.0015	0.8700 ± 0.0010	0.9560 ± 0.0377
	Naive Bayes	0.9318 ± 0.0009	0.9202 ± 0.0014	0.8436 ± 0.0024	0.9620 ± 0.0316
	Decision Tree	0.9448 ± 0.0009	0.9224 ± 0.0030	0.8838 ± 0.0008	0.9700 ± 0.0241
	Random Forest	0.9507 ± 0.0007	0.9424 ± 0.001	0.8848 ± 0.0018	0.9720 ± 0.0204
	Gradient Boosting	**0.9540 ± 0.0008**	0.9430 ± 0.0017	**0.8909 ± 0.0013**	0.9780 ± 0.0166
	AdaBoost	0.9516 ± 0.0007	**0.9431 ± 0.0013**	0.8840 ± 0.0010	**0.9800 ± 0.0179**
Phen. Aware	LR (L1)	0.8635 ± 0.0293	0.8694 ± 0.0254	0.8030 ± 0.0223	0.9700 ± 0.0257
	LR (L2)	0.8632 ± 0.0331	0.8771 ± 0.0262	0.8005 ± 0.0252	0.9580 ± 0.0166
	Naive Bayes	0.8691 ± 0.0290	0.8731 ± 0.0232	0.7659 ± 0.0509	0.9460 ± 0.0156
	Decision Tree	0.8853 ± 0.0261	0.8698 ± 0.0257	0.8179 ± 0.0133	0.9340 ± 0.0559
	Random Forest	**0.8953 ± 0.0234**	0.8833 ± 0.0213	**0.8314 ± 0.0101**	0.8340 ± 0.2057
	Gradient Boosting	0.8924 ± 0.0241	**0.8856 ± 0.0232**	0.8204 ± 0.0116	0.9440 ± 0.0280
	AdaBoost	0.8704 ± 0.0273	0.8651 ± 0.0259	0.8024 ± 0.0189	**0.9780 ± 0.0140**

Table 1: Pair prediction results. Values are means and standard deviations aggregated over ten test sets. "Random" refers to random hold-out and "Phen. Aware" to phenotype-aware hold-out. AUROC: area under the receiver operating curve, AUPRC: area under the precision-recall curve, Max. F1: maximum F1 score, Prec.@50%: precision at k=50% of the positive test cases.

Our models generally perform better in the random hold-out setting compared to the phenotype-aware hold-out setting. Note that in both settings ensemble methods (random forest, gradient boosting, and adaboost) perform best, though simpler models are often competitive.

Results for the various subset models

For L1-penalized logistic regression, the most positively predictive covariates were the "Lin measure coefficient" and "information coefficient", with beta coefficients 7.290 and 8.868 respectively. An ancestor relationship between the concepts and same visit co-occurrence were the most negatively weighted features (-3.059). We found that the semantic similarity features contributed the most to our performance, with an AUROC of 0.94 and AUPRC of 0.74 on its own. This was followed by lexical features, which on their own had an AUROC of 0.78 and AUPRC of 0.44. The concept co-occurrence and concept embedding had low performance of AUROC 0.67.

Discussion and Conclusions

Our results suggest that our models are capable of predicting which concept pairs should belong to a phenotype concept set. In addition, the most informative features for positive concept pair prediction are semantic features derived from ontologic concept distance, with some additional information derived from lexical features. Our models also performed better with a random hold-out test set likely due to the presence of concept pairs which provide context for the pairs we aim to recover in our test set. In future work we will develop methods to propose candidate concept sets for novel phenotype definitions given concept pair predictions produced by our models.

References

1. Hripcsak G, Albers DJ. High-fidelity phenotyping: richness and freedom from bias. *Journal of the American Medical Informatics Association* . 2017.
2. Gottesman,O.,Kuivaniemi,H.,Tromp, G.,Faucett, W.A.,Li, R., Manolio, Teri A, et. al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genetics in Medicine.* 2016.
3. Kirby, Jacqueline C., et al.PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. Journal of the American Medical Informatics Association 2016.
4. Jiang X, Kalluri KSD, Lee J, Liu C, Pang C, Natarajan, K, Ryan, P. Feature Engineering to Power Machine Learning for Phenotype Development. OHDSI Symposium 2019 (prospective).