SAND2020-1765C

Sandia National Laboratories

Causal inference modeling aids in feature selection for QSAR machine learning models

Bernard Nguyen, Leanne Whitmore, Anthe George, Corey M. Hudson
Sandia National Laboratories – Livermore, CA

Introduction:

Evaluating candidate biofuels can often be an expensive process, from the synthesis and production of the fuel to specialized engine testing and environmental screening. As a result, in-silico screening of molecules based on chemical and fuel properties is an essential step in the biofuels research process. Chemical kinetics simulations are computationally expensive, often taking thousands of computer hours to screen a single compound. Existing machine learning **(ML)** models are faster, but due to limited data are often inaccurate and imprecise. Here we propose better feature selection via causal inference modeling in order to improve existing ML models for various fuel properties.

Molecular descriptors are often used in quantitative structure-activity relationship **(QSAR)** models as digital representations of molecular structure. The infinite number of possible permutations of elemental composition and intramolecular connectivity results in an overwhelmingly vast feature space, and as a result, slow training times and poor predictive performance. As shown in Figure 1 below, these descriptors can often be very abstract or very specific, with any given set of molecular descriptors containing hundreds or thousands of possible substructures.

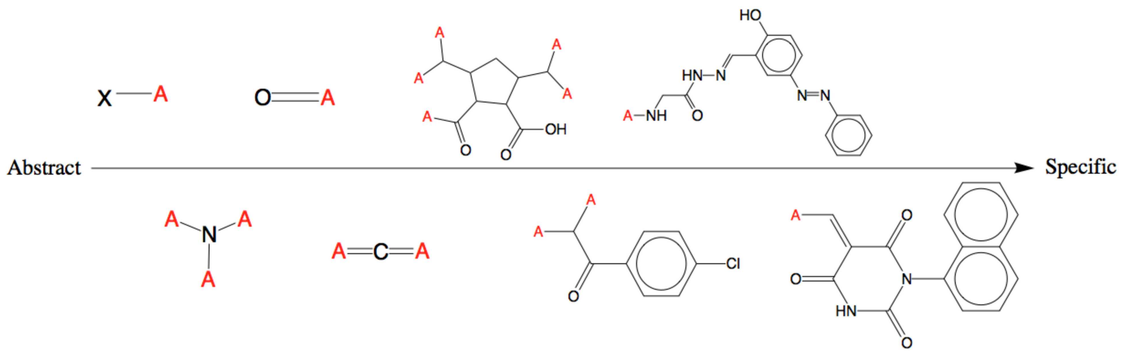

Figure 1: Examples of molecular descriptors

The existing binary classification models for research octane number **(RON)**, motor octane number **(MON)**, and octane sensitivity **(OS)** utilize these molecular descriptors and are good at predicting whether a molecule will have high or low RON/MON/OS, but fail to predict precise and accurate metrics. Our hypothesis is that we can improve model performance by selectively including only **relevant** and **important** features. In other words, **"which molecular substructures cause a given fuel metric to go up or down?"**

Methods:

In order to evaluate causal inference **(CI)** as a means for feature reduction, we compared it to other common methods for evaluating feature importance:

- Spearman correlation coefficient
- Euclidean distance (when feature is included vs omitted)
- Gini importance (provided by the random forest model)
- Local Interpretable Model-agnostic Explainer **(LIME)**[2]

Causal effects were estimated using **DoWhy**[3], a python library for causal inference based on Judea Pearl's do-calculus. Specifically, the linear regression **(CILR)** and propensity score stratification **(CIPSS)** estimators were used for this study.

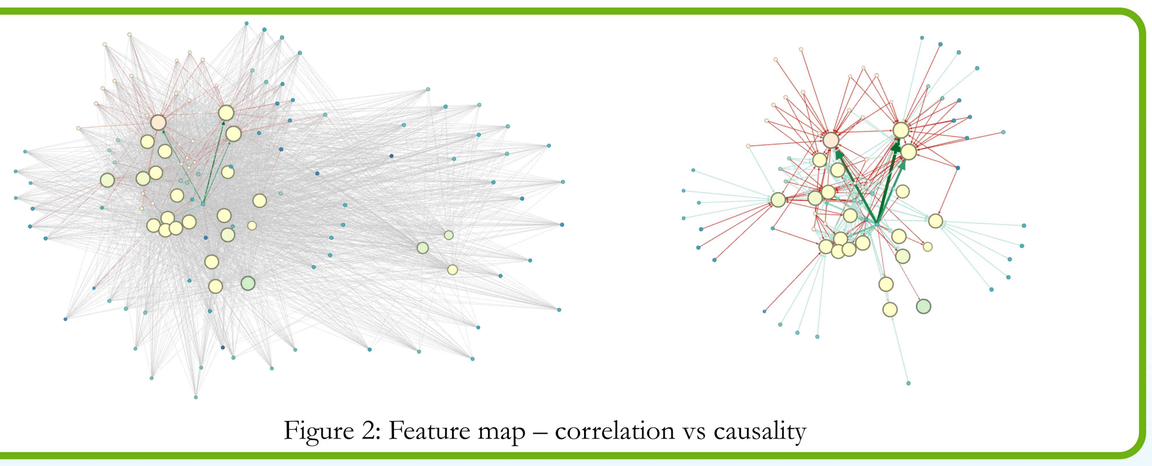

Figure 2: Feature map – correlation vs causality

Each metric produced a different ranking of feature importances. Using these rankings, we added features one-by-one to the machine learning models and quantified model performance using the root mean squared error **(RMSE).** The regression models used for comparison in this experiment were XGBoost regressors with fairly standard parameters (100 estimators, max depth of 3).

Results:

Ranking Method	RMSE Range			# Features (80% Weight)		
	RON	**MON**	**OS**	**RON**	**MON**	**OS**
CILR	15.48 - **6.31**	13.59 - 6.27	7.84 - **4.50**	237	213	210
CIPSS	11.67 - 7.84	**10.63** - 6.01	**7.63** - 4.69	28	34	40
Euclidean	16.97 - 7.40	17.70 - 6.95	8.96 - 4.60	334	335	338
Gini	17.42 - 8.19	16.90 - 6.75	8.11 - 4.97	44	43	34
LIME	16.47 - 8.10	14.94 - 6.68	8.03 - 4.66	8	6	6
Spearman	**11.24** - 7.97	12.84 - **5.94**	8.59 - 4.53	217	212	208

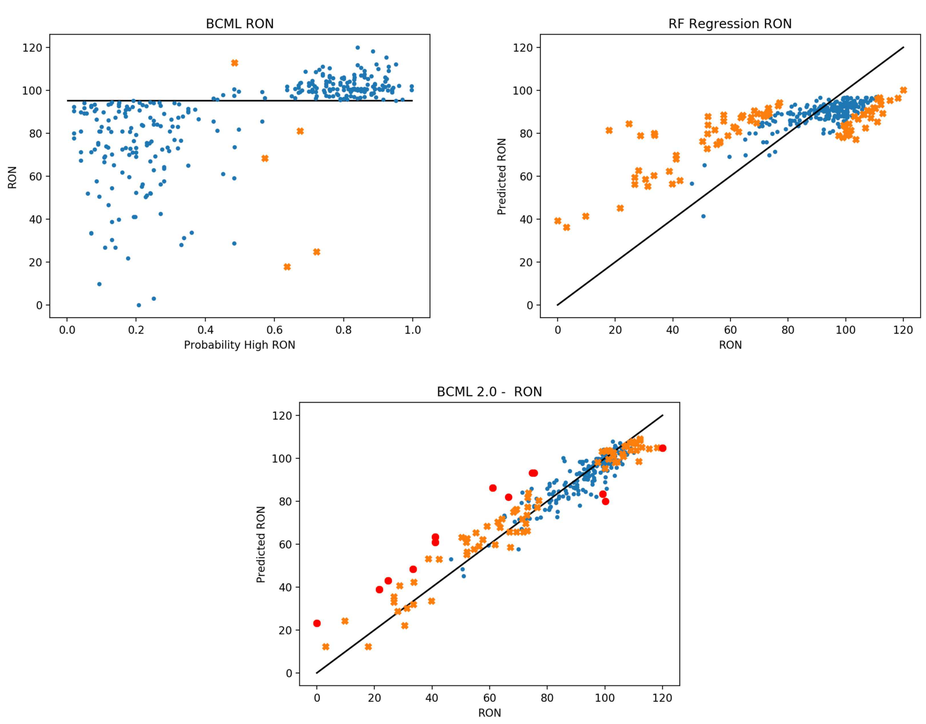

Figure 3: Original BiocompoundML (RON) classifier and regressors with and without filter reduction (CIPSS).

Conclusions:

The generally low RMSE when few features are considered (upper bound) tells us that CIPSS and LIME are some of the best metrics at identifying the ***most*** important features, but the failure to reach the minimum RMSE with all features of non-zero importance tells us that they fail to consider ***all*** relevant features. Model results with and without feature reduction (Figure 3) clearly demonstrate the importance of feature reduction as a preliminary step for all QSAR models. That being said, it is important to note that feature selection is by no means an alternative for hypertuning of ML model parameters.

The new ML models for RON and MON have an approximate error of ±10 in our test set and can now be used to tentatively screen molecules in-silico.

References:

[1] Whitmore, Leanne S., et al. "BioCompoundML: A General Biofuel Property Screening Tool for Biological Molecules Using Random Forest Classifiers." *Energy & Fuels,* vol. 30, no. 10, 2016, pp. 8410-8418., doi:10.1021/acs.energyfuels.6b01952.

[2] Ribeiro, Marco, et al. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2016.

[3] Sharma, Amit. "ACM KDD 2018 International Conference on Knowledge Discovery and Data Mining", *Tutorial on Causal Inference and Counterfactual Reasoning*, https://causalinference.gitlab.io/kdd-tutorial/

This research was conducted as part of the Co-Optimization of Fuels & Engines (Co-Optima) project sponsored by the U.S. Department of Energy (DOE) Office of Energy Efficiency and Renewable Energy (EERE), Bioenergy Technologies and Vehicle Technologies Office.

Co-Optimization of Fuels & Engines

Sandia National Laboratories