Automated Identification of Metal-Organic Framework Synthesis Information

David Venator, Elijah Kellner, Xintong Zhao, Jane Greenberg

Metadata Research Center (MRC), College of Computing and Informatics, Drexel University, Philadelphia, PA

Motivation

Reporting of chemical synthesis methods in research literature is often incomplete. This is especially true with studies concerning Metal-Organic Frameworks (MOFs). As a result, many scientists in the field have limited idea how to synthesize or replicate desired MOF structures.

Goals and Objectives

The goal of this study is to use machine learning to extract synthesis data to aid researchers seeking to synthesize these materials.

Specific objectives are to:

- Build training dataset based on existing MOFs literature for the binary classification task (179)
- Generate features from text data for machine learning models
- Test decision tree, random forest, supporting vector machine, and logistic regression machine learning models and analyze their performance

Methods and Process Pipeline

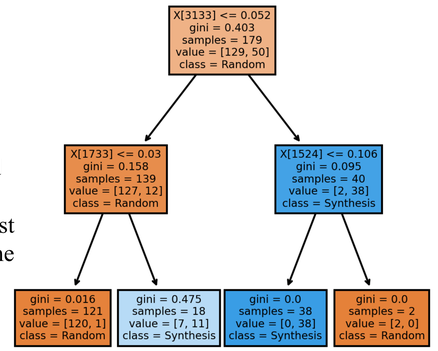
- Download metal organic framework related research journal papers
- Annotation: We annotate paragraphs describing synthesis procedure as positive example, and we randomly select other paragraphs from articles as negative example. The percentage of positive examples in the dataset is roughly 33%.
- We applied TF-IDF, Bag-of-Word models to extract features from text data
- SVM, Random Forest, logistic regression models were used
- We use precision, recall and f1-score as evaluation metrics

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \qquad Recall = \frac{True\ Postive}{True\ Positive + False\ Negative}$$

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Decision Tree/Random Forest

- This tree is the result of a random forest model constructed using TF-IDF extracted features
- Key positive indicators are mmol (3133), co2 (1524), and crystals (1733).
- Results show the random forest has better performance than the decision tree algorithm.

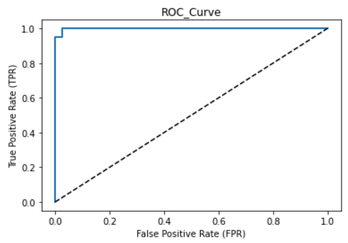


Decision Tree

	precision	recall	f1-score	support
0	0.95	0.93	0.94	40
1	0.86	0.90	0.88	20

Random Forest

	precision	recall	f1-score	support
0	0.95	1.00	0.98	40
1	1.00	0.90	0.95	20

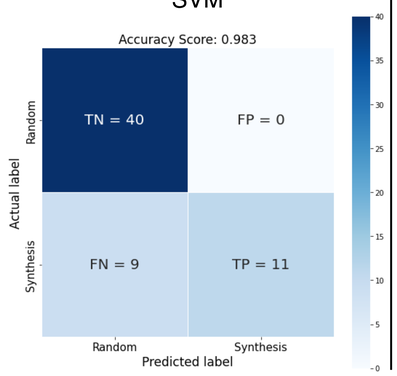


Logistic Regression/SVM

- Classification models for SVM and Logistic Regression were also created – logistic regression being very high performing while SVM had very low recall

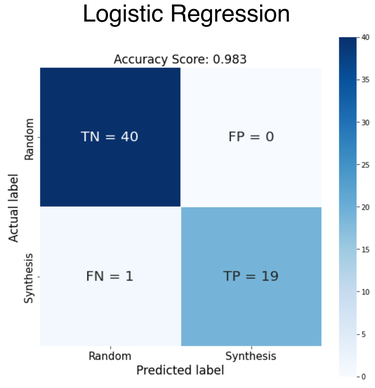
	precision	recall	f1-score	support
0	0.82	1.00	0.90	40
1	1.00	0.55	0.71	20

SVM



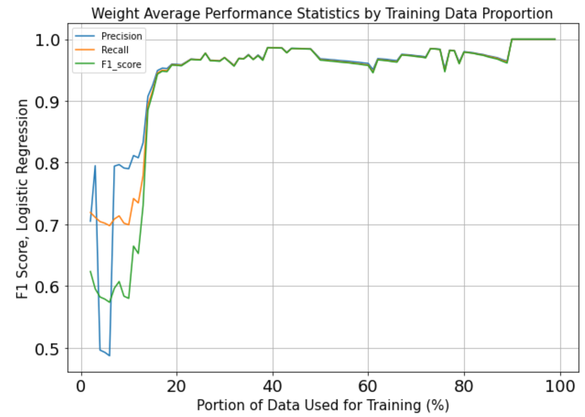
	precision	recall	f1-score	support
0	0.98	1.00	0.99	40
1	1.00	0.95	0.97	20

Logistic Regression



Performance Evaluation

The highest performing model, logistic regression, performs to a similar standard with as little as 20% training data



Conclusions / Future Work

- Successfully collected and annotated large dataset of MOF synthesis – processed textual data using python tools
- Created and evaluated machine learning models to identify synthesis paragraphs
- Determined classification capabilities of highest performing model (logistic regression)

- Develop flowchart methodology to understand and visualize MOF synthesis procedures

Acknowledgments

We acknowledge support of NSF-HDR-OAC #2118201 Institute for Data Driven Dynamical Design. We also acknowledge REU infrastructure support via NSF-EEC-ENG #1949718 Smart Manufacturing Research Experiences for Undergraduates (SMREU). We would like to thank Dr. Jane Greenberg, Xintong Zhao, and Scott McClellan for their support and guidance during this project.