Application of Machine Learning to Materials Discovery and Development

Ankit Agrawal and Alok Choudhary
Department of Electrical Engineering and Computer Science
Northwestern University
{ankitag,choudhar}@eecs.northwestern.edu

Contributors:
Surya Kalidindi (GaTech), Basavarsu (TRDDC), Chris Wolverton (NU),
Ahmet Cecen (GaTech), Parijat Deshpande (TRDDC), Bryce Meredig (NU)

MURI 3-Year Review
June 22-23, 2015
Integrated Computational Materials Engineering (ICME)

Project Collaboration

Project I. Multi-objective Structure-Property Optimization

Project Collaboration

**Project I.** Multi-objective Structure-Property Optimization

**Project II.** Multiscale Prediction of Localization Relationships

Project Collaboration

Project I. Multi-objective Structure-Property Optimization

Project II. Multiscale Prediction of Localization Relationships

Project III. Exploring Composition-Processing-Property Relationships

Project IV. Composition-based Discovery of Stable Compounds

Objective: Employ data-driven approaches to the NIMS public domain materials database for exploring composition-processing-property relationships and constructing predictive models for fatigue strength of steels.
## NIMS Database Attributes

### Fatigue Data Sheet Information:

**Chemical composition** - %C, %Si, %Mn, %P, %S, %Ni, %Cr, Cu %, Mo% (all in wt. %)

**Upstream processing details** - Ingot size, Reduction ratio, Non-metallic inclusions

**Heat treatment conditions** – Temperature, Time and other process conditions for Normalizing, Carburizing-Quenching and Tempering processes

**Mechanical properties** - YS, UTS, %EL (Elongation), %RA (Reduction in Area), Vickers Hardness, Charpy impact value (J/cm²), Rotating bending fatigue strength @ $10^7$ cycles

Total - 437 data records
- Carbon and low alloy steels - 371 observations
- Carburizing steels - 48 observations
- Spring steels - 18 observations

Steel Fatigue Strength Prediction Framework

1. Preprocessing
2. Feature Selection
3. Predictive modeling
4. Evaluation

- Rotating Bending Fatigue Testing Data from NIMS
- Preprocessed Data
- Fatigue Strength Prediction Database
- Training Split

Testing Split

Leave One Out Cross Validation (LOOCV)
Cluster Visualization
Information Gain Based Feature Ranking
Evaluation Metrics

- Compare vectors of actual and predicted values
  - Coefficient of correlation (R)
  - Coefficient of determination (R²)
  - Mean Absolute Error (MAE)
  - Root Mean Squared Error (RMSE)
  - Standard Deviation of Error (SDE)
  - Mean Absolute Error Fraction (MAE)
  - Root Mean Squared Error Fraction (RMSE)
  - Standard Deviation of Error Fraction (SDE)

\[
R = \frac{\sum_{i=1}^{N} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2 \sum_{i=1}^{N} (\hat{y}_i - \bar{\hat{y}})^2}}
\]

\[
MAE = \bar{e} = \frac{1}{N} \sum_{N} |y - \hat{y}|
\]

\[
RMSE = \sqrt{\frac{1}{N} \sum_{N} (y - \hat{y})^2}
\]

\[
SDE = \sqrt{\frac{1}{N} \sum_{N} (|y - \hat{y}| - \bar{e})^2}
\]

\[
MAE_f = \bar{e}_f = \frac{1}{N} \sum_{N} \left| \frac{y - \hat{y}}{y} \right|
\]

\[
RMSE_f = \sqrt{\frac{1}{N} \sum_{N} \left( \frac{y - \hat{y}}{y} \right)^2}
\]

\[
SDE_f = \sqrt{\frac{1}{N} \sum_{N} \left( |\frac{y - \hat{y}}{y} - \bar{e}_f| \right)^2}
\]
Results Comparison

- R^2
- MAE_f
- RMSE_f
- SDE_f

Legend:
- DecisionTable
- LRTrans
- ANN
- Instance-based
- RobustFitLSR
- REPTree
- KStar
- LinearRegression
- M5ModelTree
- SVM
- PaceRegression
- MPR
## Results Comparison

<table>
<thead>
<tr>
<th>Method</th>
<th>$R$</th>
<th>$R^2$</th>
<th>MAE</th>
<th>RMSE</th>
<th>SDE</th>
<th>MAE$_f$</th>
<th>RMSE$_f$</th>
<th>SDE$_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>DecisionTable</td>
<td>0.9494</td>
<td>0.9014</td>
<td>34.8762</td>
<td>58.5932</td>
<td>47.1371</td>
<td>0.0584</td>
<td>0.0806</td>
<td>0.0557</td>
</tr>
<tr>
<td>IBk</td>
<td>0.9589</td>
<td>0.9195</td>
<td>46.0320</td>
<td>53.2749</td>
<td>26.8499</td>
<td>0.0859</td>
<td>0.0940</td>
<td>0.0382</td>
</tr>
</tbody>
</table>

Research

**Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters**

Ankit Agrawal$^{1*}$, Parijat D Deshpande$^2$, Ahmet Cecen$^3$, Gautham P Basavarsu$^2$, Alok N Choudhary$^1$ and Surya R Kalidindi$^{34}$

<table>
<thead>
<tr>
<th>Method</th>
<th>$R$</th>
<th>$R^2$</th>
<th>MAE</th>
<th>RMSE</th>
<th>SDE</th>
<th>MAE$_f$</th>
<th>RMSE$_f$</th>
<th>SDE$_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>REPTree</td>
<td>0.9862</td>
<td>0.9726</td>
<td>22.5671</td>
<td>30.9401</td>
<td>21.1907</td>
<td>0.0414</td>
<td>0.0542</td>
<td>0.0349</td>
</tr>
<tr>
<td>M5ModelTree</td>
<td>0.9890</td>
<td>0.9781</td>
<td>19.3760</td>
<td>27.6065</td>
<td>19.6870</td>
<td>0.0353</td>
<td>0.0484</td>
<td>0.0332</td>
</tr>
<tr>
<td>MPR</td>
<td>0.9900</td>
<td>0.9801</td>
<td>18.5529</td>
<td>26.4378</td>
<td>18.8563</td>
<td>0.0350</td>
<td>0.0556</td>
<td>0.0432</td>
</tr>
</tbody>
</table>

Discovery of stable compounds

Collaborative project between Agrawal (NU), Choudhary (NU), Wolverton (NU)
Discovery Framework

(a) Discovery Framework Overview
- **Database Construction**
  - Thousands of DFT formation energies
  - Empirical elemental data
- **Predictive Modeling**
  - Model 1: established heuristic
  - Model 2: data mining
- **Model Evaluation**
  - Test models on unseen formation energies

(b) Detailed Process
- **Prediction**
  - Run combinatorial list of compositions through models
- **Ranking**
  - Combine heuristic and data mining predictions
- **Validation**
  - Experiments
  - Crystal structure prediction

- Millions of candidate ternary compositions
- Models
- Formation energy predictions
- Ranked high-potential candidates
- Compound discovery
Model Validation: Ranking

Combined model outperforms either alone in regime of interest
What happens when we rank “all possible ternaries” by their likelihood of stability?
Predictions for Discovery

Average of all A-B-X ternaries

Fingerprint of entire unexplored ternary composition space!

Interesting insights:

- Highest ranked ternary: $SiYb_3F_5$
  - $Si$ acts as an anion
  - Validated with structure and DFT calculations
- pnictides, chalcogenides, halides
- $Pt-X-Y$
- $Ti_{12}SrTe_{19}$ – a missing binary $Ti_2Te_3$?
Example of discovered stable ternary compositions whose stability was explicitly confirmed with crystal structure prediction. Our method is successful at identifying new stable compounds across a wide variety of chemistries.

Summary

➢ Steel Fatigue Strength Prediction
  o NIMS database consisting of composition and processing parameters linked with performance (fatigue strength).
  o Neural networks, decision trees, multivariate polynomial regression able to achieve high $R^2$ values of >0.98.

➢ Stable Compound Discovery
  o A database of DFT calculations used to learn composition-property relationships, thus mimicking DFT for estimating stability.
  o The resulting predictive models used to scan the entire ternary composition space to discover likely stable compositions.
  o Many predictions explicitly confirmed with crystal structure prediction and DFT.
Future Outlook

Project I. Multi-objective Structure-Property Optimization

Project II. Multiscale Prediction of Localization Relationships

Project III. Exploring Composition-Processing-Property Relationships

Project IV. Composition-based Discovery of Stable Compounds


Thank You!

Ankit Agrawal
Research Associate Professor
Dept. of Electrical Engineering and Computer Science
Northwestern University
ankitag@eecs.northwestern.edu
www.eecs.northwestern.edu/~ankitag/