**Citrine Data Science Challenge Report – Predicting Stable Binary Compounds**

Fang Ren

**Background**

The thermodynamic stability of a compound concerns how readily the chemical bonds break down. The formation of enthalpy () can be used to characterize this property. If of a compound is large and negative, some large amount of energy must be applied to the compound to break its chemical bonds, and thus it is thermally stable. If is positive, the compound is unstable. A simple threshold can be used to represent this information, as shown in Table 1. Note that whether a compound is stable or not sometimes also depends on the entropy change when it decomposes. For simplicity, this is not included in the current prediction.

Table 1. Connection of formation of enthalpy and thermodynamic stability of a compound

|  |  |  |
| --- | --- | --- |
| Formation of enthalpy | thermodynamic stability | Stability label |
|  | Unstable | 0 |
|  | Stable | 1 |

The prediction of the thermodynamic stability of a compound can then be performed by first predicting . After that, the threshold in Table 1 can be applied to classify whether the compound is thermally stable or not. This has been a popular topic in materials science, and DFT1-3, CSP4, machine learning5-6 methods and some other theoretical models7-8 have all been used to tackle the problem. Materials database9 has also been built to provide quantitative knowledge of composition-formation energy relationship. Because is a continuous value, all of these methods fall into a regression category.

It is therefore a novel approach to solve the same problem using classification, by using training data that are already labeled as “1-stable” or “0-unstable”. However, it is possible that the classification may not perform as well as the first one, because significant amount of information is discarded in the training data (for example, a compound can be somewhat stable, but another may be extremely stable, but both of them will be labeled as “1”). Another reason that classification may not work as well as regression is that, the stability of compounds change gradually, and thus there in fact do not exist two distinctive classes that can effectively separate them. Therefore, classification is likely to generate many false positive and negatives. On the other hand, labeled training data does provide a cleaner classification trees, and it would be interesting to compare the predicted results from classification with regression. In this report, I will use several models based on either random forest classifier10 or neural network classifier11, to predict the stability of unknown compounds, and assess the performance of all of these models.

**Results and Discussions**

Organize the training data

The training data has 2572 binaries, and each binary has 96 features (The element formula and atomic number are redundant, so only one of them is used). The number of samples is relative small, and it is not likely to produce good predicted results. A visualization of the input layer and output layer is shown in Figure 1.

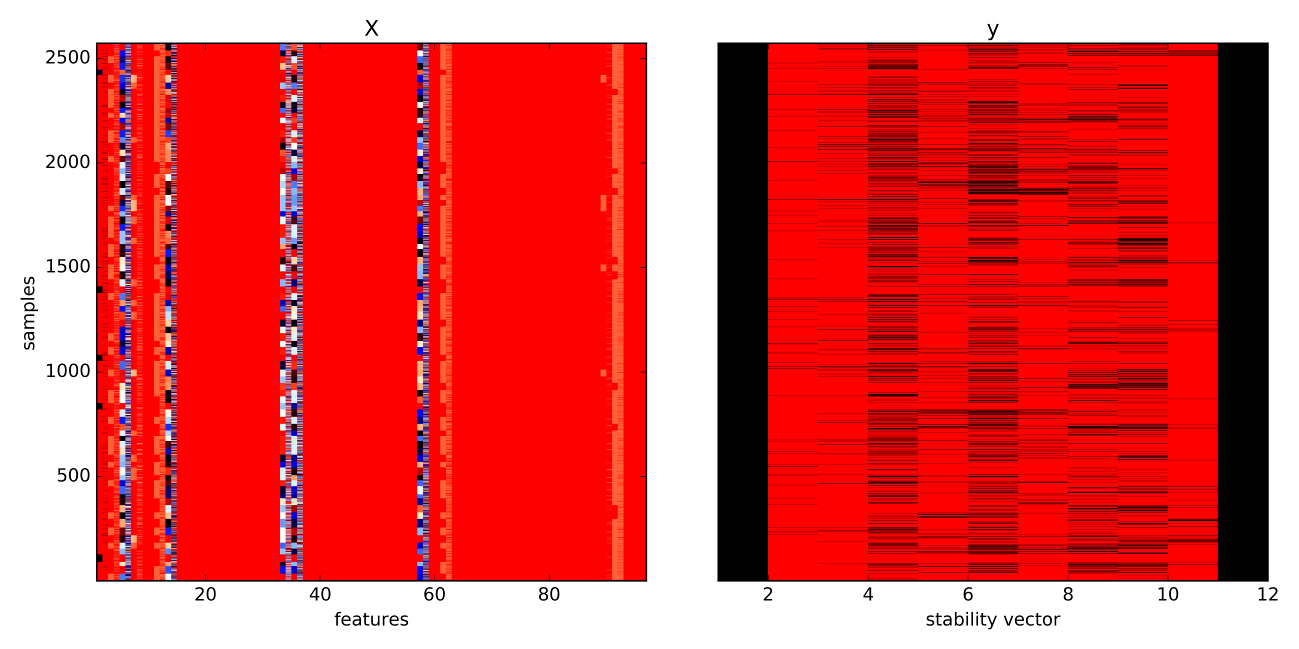


Figure 1. Training data input layer X and output layer y.

It is important to notice that the output layer “y” has 11 dimensions, which correspond to 11 distinctive compositions at a 10% interval (e.g., for element A and B, there will be A, A0.9B0.1, A0.8B0.2, A0.7B0.3, A0.6B0.4, A0.5B0.5, A0.4B0.6, A0.3B0.7, A0.2B0.8, A0.1B0.9, and B). Based on this observation, there are potentially two approaches to use the training data. 1. Use the input layer as it is, and there are 11 independent output layers, each with a distinctive composition (for example A0.7B0.3 and C0.7D0.3 are in the same output layer, while A0.7B0.3 and C0.6D0.4 are not). 2. Combine the original input layer and an additional composition feature (ranged from 0 to 10), into a new input layer (97 features), and the output layer will be single layered.

The main difference of the two approaches is whether two binaries having different compositions (e.g. A0.7B0.3 and C0.6D0.4) should have influences on each other, and the answer is probably yes. The second approach will also increase the training sample number from 2572 to 28,292(2572×11), which is more likely to produce more precise results. The visualization after re-organizing the training data is shown in Figure 2. In this report, I will compare the results using both of the approaches and validate my hypothesis.

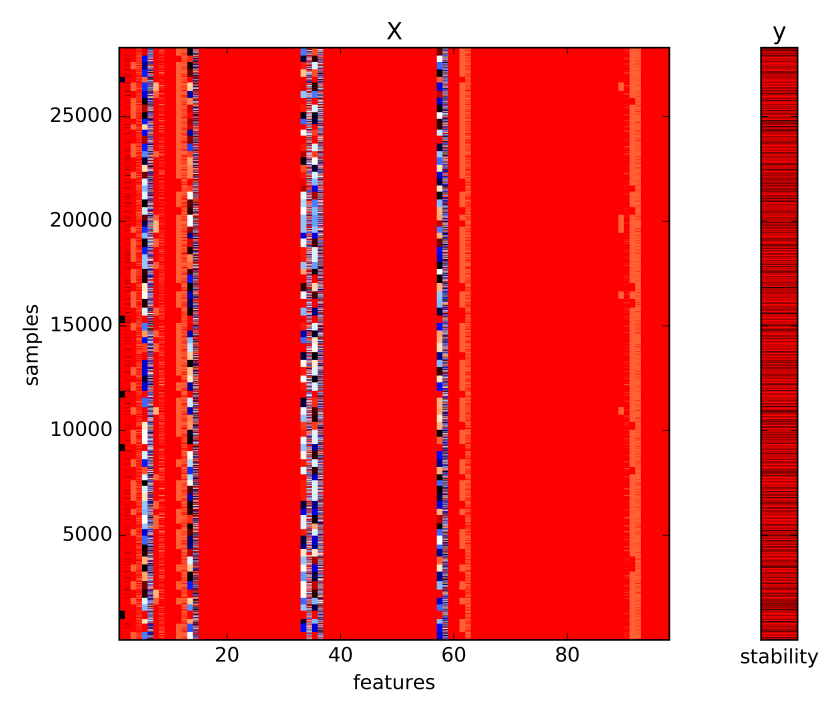


Figure 2. Training data input layer X and output layer y after reorganization.

Model selection - random forest

The prediction results on stability of a binary using random forest are shown below, using original training data (Figure 3) and reorganized training data (Figure 4). The first 2000 samples (for reorganized data, it would be the first 22,000 samples) were included as training data, and the rest 572 samples (for reorganized data, it will be 6,292 samples) were used to validate the prediction. Notice that the f-measures for both of the models are not very good (0.05 and 0.4), the model using re-organized training data still have a much better performance. In both of the models, the pure elements were not included for training or for f-measure calculation, because they do not add much information to model and are likely to produce false high f-measures, since the prediction on pure elements will always be precise (f-measure = 1).

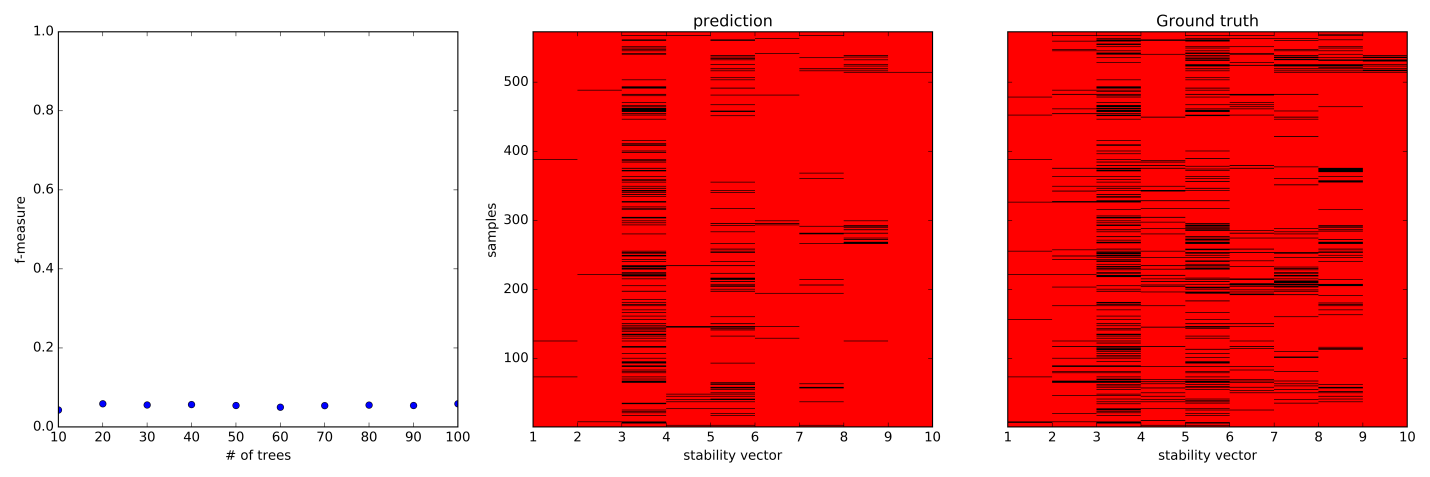


Figure 3. Random forest (20 trees) classification results on binary stability prediction without reorganizing input.

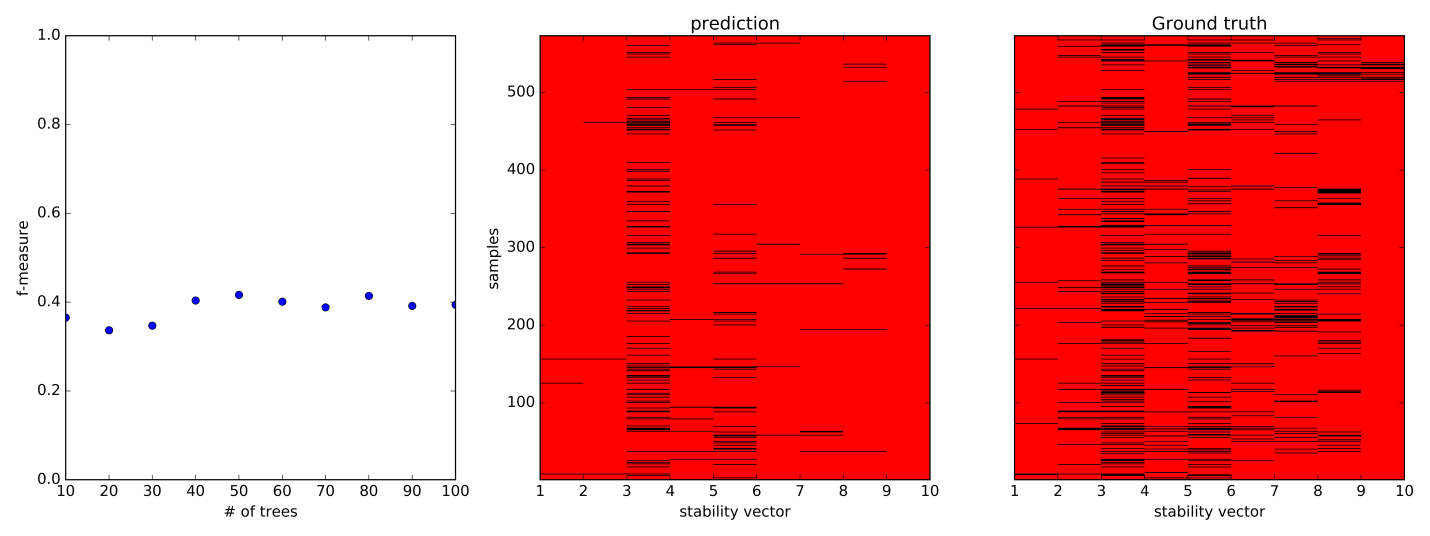


Figure 4. Random forest (40 trees) classification results on binary stability prediction without reorganized input.

Model selection - neural network

Next, several neural network models were built on the same training data, and the performance will be assessed and compared with that predicted by random forest models. A brief discussion on choosing different parameters of neural network architecture is as follows. A common empirical rule for choosing hidden layers is the optimal size of the hidden layer is usually between the size of the input and size of the output layers. The number input neurons in this dataset is 96 for original training data and 97 for re-organized training data. The number of neurons in the output layer is 1. Another empirical rule for hidden layers is that the number of neurons decreases monotonically when approaching the output layer12. For Activation function, because the target is a binary (0/1), “logistic function” is chosen13 to compare with the default rectified linear unit function. The ‘lbfgs’ function is also considered as the solver function because it is supposed to converge faster and perform better for a small dataset11. All of the other parameters were set as default.

Table 2. Neural network model configuration.

|  |  |  |  |
| --- | --- | --- | --- |
| MLP | configurations | Activation function | Solver function |
| mlp1 | 48 | “relu” or “logistic” | “lbfgs” or “adam” |
| mlp3 | 96-48-1 |
| mlp5 | 96x2-48-24-1 |
| mlp7 | 96x2-48x2-24x2-1 |
| mlp9 | 96x2-48x2-24x2-12x2-1 |

The f-measures were calculated by alternating the hidden-layer configuration, the activation function, and the solver. However, none of these models produced reasonable results, even though the the models with the highest f-measure are shown in **Figure 5** (without reorganizing input) and Figure 6 (with reorganized input. The reasons why the models failed may include: 1. The number of training records is not large enough. For example, this paper6 has a collection of over 285,780 entries as training data for formation energy. 2. This is not a good classification problem because the stability of a compound is a spectrum instead of binary.

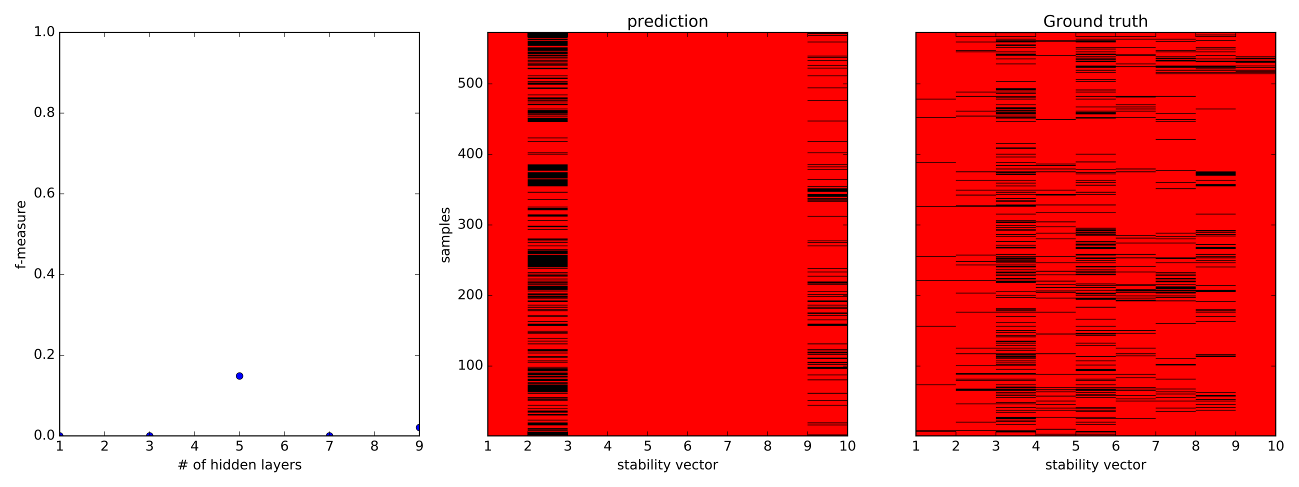


Figure 5. Neural network (5 hidden layers, activation = “relu”, solver = “lbfgs”) classification results on binary stability prediction without reorganizing input.

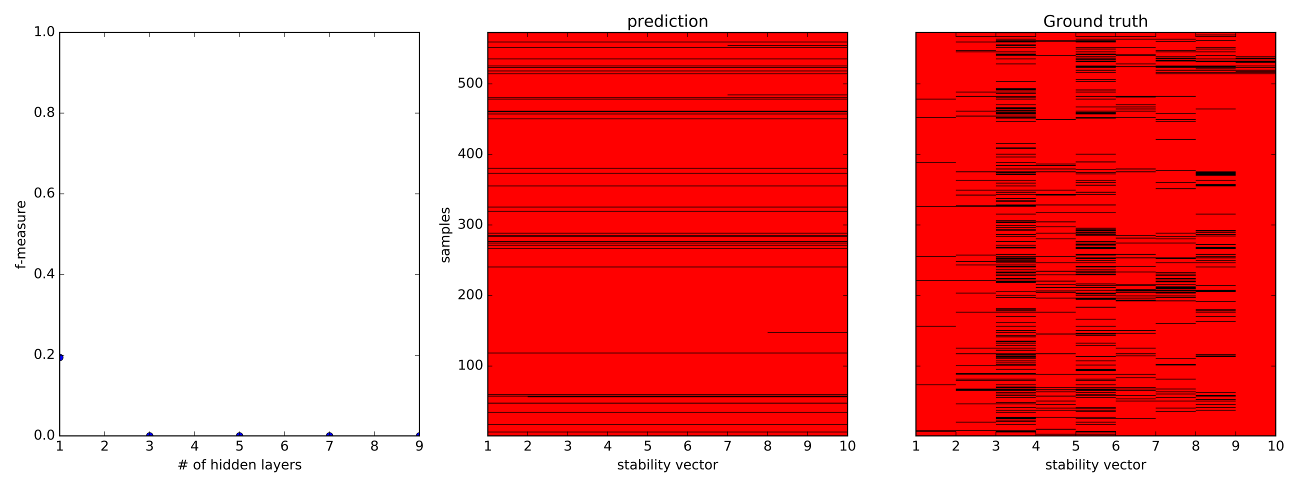


Figure 6. Neural network (1 hidden layer, activation = “relu”, solver = “adam”) classification results on binary stability prediction with reorganized input.

In summary, I have compared a few models based on random forest and neural network. Even with the best model presented here, the f-measure is only about 0.4, which means more than half of the time, the model will predict the stability of a compound wrong. The reason for that is the stability of a chemical compound is a spectrum, apart from some extremely stable compounds like SiO2 or extremely unstable compounds like N2O4, most of the compounds lie in between. Simply dividing them into two groups will likely lose significant amount of information and is unlikely to produce reasonable result.

**References**

1. Stevanovic, V.; Lany, S.; Zhang, X. W.; Zunger, A., Correcting density functional theory for accurate predictions of compound enthalpies of formation: Fitted elemental-phase reference energies. *Phys Rev B* **2012,** *85* (11).

2. Colinet, C., Ab-initio calculation of enthalpies of formation of intermetallic compounds and enthalpies of mixing of solid solutions. *Intermetallics* **2003,** *11* (11-12), 1095-1102.

3. Mole, S. J.; Zhou, X. F.; Liu, R. F., Density functional theory (DFT) study of enthalpy of formation .1. Consistency of DFT energies and atom equivalents for converting DFT energies into enthalpies of formation. *J Phys Chem-Us* **1996,** *100* (35), 14665-14671.

4. Zhu, Q.; Oganov, A. R.; Zhou, X. F., Crystal Structure Prediction and Its Application in Earth and Materials Sciences. *Top Curr Chem* **2014,** *345*, 223-256.

5. Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C., Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys Rev B* **2014,** *89* (9).

6. R Liu, L. W., A Agrawal, W Liao, C Wolverton, A Choudhary, Deep Learning for Chemical Compound Stability Prediction. In *KDD Workshop on Large-Scale Deep Learning for Data Mining*, San Francisco, CA, 2016.

7. Muthurajan, H.; Sivabalan, R.; Talawar, M. B.; Anniyappan, M.; Venugopalan, S., Prediction of heat of formation and related parameters of high energy materials. *J Hazard Mater* **2006,** *133* (1-3), 30-45.

8. Mostafa, A. T. M. G.; Eakman, J. M.; Yarbro, S. L., Prediction of Standard Heats and Gibbs Free-Energies of Formation of Solid Inorganic Salts from Group Contributions. *Ind Eng Chem Res* **1995,** *34* (12), 4577-4582.

9. Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C., Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *Jom-Us* **2013,** *65* (11), 1501-1509.

10. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P., Random forest: A classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comp Sci* **2003,** *43* (6), 1947-1958.

11. Neural network models (supervised). <http://scikit-learn.org/stable/modules/neural_networks_supervised.html>.

12. <http://www.faqs.org/faqs/ai-faq/neural-nets/part1/preamble.html>.

13. Jordan, M. I. *Why the logistic function? A tutorial discussion on probabilities and neural networks*; Massachusetts Institute of Technology: Computational Cognitive Science Technical Report 9503, 1995; p 1.