Optimizing Positive-Unlabeled Learning Method in Inorganic Material Synthesizability Prediction Through Hyperparameter Tuning

William Cai
Stanford University

I. Nomenclature

PU Learning = Positive-Unlabeled Learning
PU Learning = Positive-Unlabeled Learning

C = Hyperparameter that trades off the correct classification of training examples against maximization of the decision function's margin

Gamma = Hyperparameter that defines how far the influence of a single training example reaches, with low values meaning 'far' and high values meaning 'close'

C-Gamma Space = 2D space contains the C and Gamma values
Positive = material is synthesizable
Unlabeled = material that is not yet synthesized in the lab
ICSD = Inorganic Crystal Structure Database
SVM = Support Vector Machine
RBF = Radial Basis Function

loss rate = 1 - (number of predicted true positive data / number of true positive data in the testing set)
pos-pred % = number of predicted positive data / number of total data in the testing set

II. Abstract

Hyperparameter tuning is critical for the performance of a machine learning model and highly depends on the problem the model is applying to. At the same time, the tuning process could be time-consuming and resource-expansive. This project attempts to optimize the Weighted Elkanoto Classifier in the inorganic material synthesizability prediction problem through hyperparameter tuning. The tuning process involves a proper sampling plan selection, surrogate model construction, and implementation of the off-the-shelf optimization method. The tuning process enables us to find the approximated optimized frontier in C-Gamma space.

III. Introduction

Traditional approaches to identify stable and synthesizable inorganic materials have largely relied on human intuition and could be computationally expensive and diversity-limited. As a result, applying data-driven approaches to search for potential new synthesizable material is becoming more popular. And among the data-driven approaches, the applications of machine learning algorithms in the material searching process are gaining more attention.

The material search process is a positive-unlabeled learning problem. This problem tackles the data set that is either being labeled as positive or is unlabeled (as a result, we do not know if the data is positive or negative). As a result, unlike the traditional machine learning process for which we can train the model with just positive and negative data set, the PU-learning method trains the model with positive data set and data set that could be either positive or negative. The PU-learning problem that this project is looking at is the inorganic material synthesizability prediction. The inorganic material whose synthesizability is experimentally verified in the labs is recorded in the ICSD and can be considered as positive labeled. On the other hand, the theoretical material that is not yet synthesized is considered unlabeled.

This project aims to build on my previous research project from the 2020 material science summer research program. In this previous research project, I implemented the supervised learning (support vector machine with RBF kernel) method to demonstrate that SVM with RBF kernel has the ability and reliability to identify materials that are already experimentally shown to be synthesizable (already in the database). However, the supervised learning methods hugely limit our ability to search and identify potential new synthesizable material that is not yet in the ICSD. As a result, I implemented the off-the-shelf method: Weighted Elkanoto Classifier to try to tackle this problem. However, Weighted Elkanoto Classifier (with its default hyperparameters) has worse performance in identifying materials that are already experimentally shown to be synthesizable (already in the database) than the supervised learning method. This problem is not yet solved in the previous summer research.

IV. Problem

The outcome of the Weighted Elkanoto Classifier is sensitive to the values chosen for tuning parameters (C and Gamma), and no good way is known to set these values. Moreover, the original application is in "identifying protein records that should be included in an incomplete specialized molecular biology database". As a result, the hyperparameters in the default package are not generalizable. The problem we are tackling is to find values of the hyperparameters so that the performance of the Weighted Elkanoto Classifier is optimized in the inorganic material synthesizability prediction problem.

V. Methods

The Sampling plan is using the uniform projection plan based on the graph of the commonly used values for C and Gamma. The code for the uniform projection plan is implemented based on the example code (Algorithm 13.2) in "Algorithms for optimization" (Mochendefer, 2019).

The surrogate model construction is using the SMT: Surrogate Modeling Toolbox.

The optimization process uses the CVXPY package.

A. Sampling Plan and Model Training

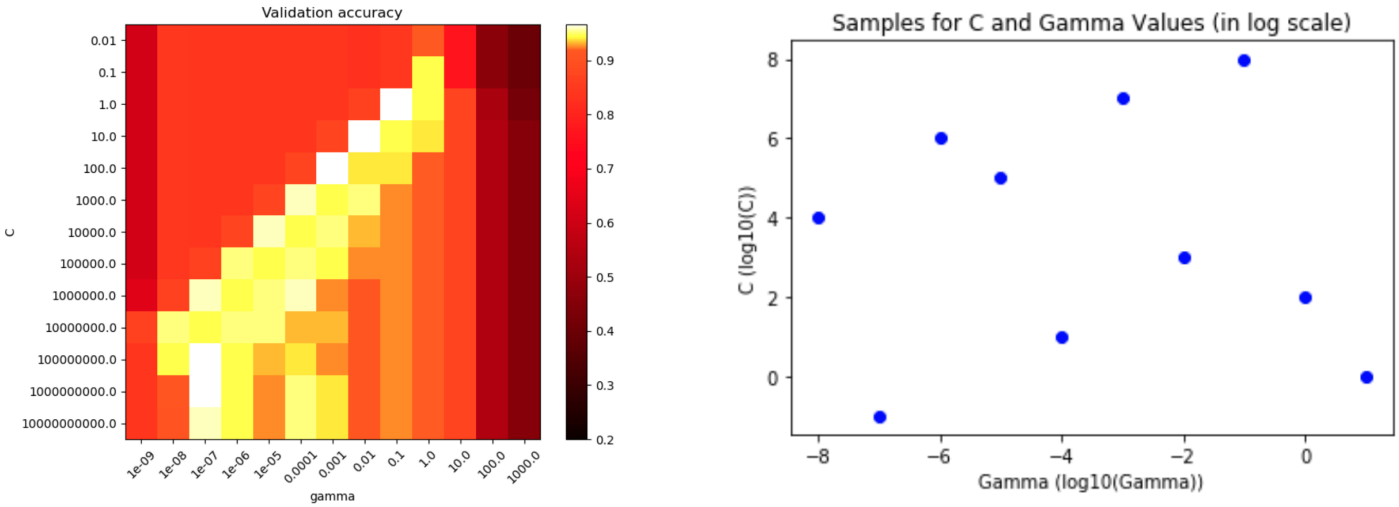

Fig. 1 Traditional C-Gamma value performance (Left) and Sampling Plan Data Visualization (Right)

The uniform projection plan for the 2D C-Gamma Space is constructed using a permutation for each dimension.

For the data used for model training, we choose the ratio of the number of positive-labeled data to the number of unlabeled data to be 1:1. In the actual implementation, all the 10899 positive labeled data from ICSD are used and 10899 unlabeled data are used.

To validate the model for each training, we use 85% of the total data for training and 15% of the total data for testing.

The loss rate and pos-pred % are the metrics used to evaluate the performance of the model (discussed in the later section).

B. Surrogate Model

After obtaining the sampling point we use these sampled hyperparameters points to obtain the loss rate and pos-pred %. Then we will use the surrogate model to approximate the function for loss rate and pos-pred %. The first surrogate model will be the value of loss rate based on the value of (C, Gamma) and the second surrogate model will be the value of *pos-pred %* based on value of (C, Gamma).

We approximate both models using the Regularized minimal-energy tensor-product splines (RMTS). RMTS is a type of surrogate model for low-dimensional problems with large datasets and where fast prediction is desired.

The prediction equation for RMTS is $y = F(x)w$.

And based on the SMT handbook, RMTS computes the coefficients of the splines, w, by solving an energy minimization problem subject to the conditions that the splines pass through the training points.

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{H}\mathbf{w} + \frac{1}{2}\beta\mathbf{w}^T\mathbf{w} + \frac{1}{2}\frac{1}{\alpha}\sum_i^{nt}\left[\mathbf{F}(\mathbf{xt}_i)\mathbf{w} - yt_i\right]^2$$

This is formulated as an unconstrained optimization problem where the objective function consists of a term containing the second derivatives of the splines, another term representing the approximation error for the training points, and another term for regularization.

C. Optimization

For loss rate surrogate function, we use cvxpy.problems.objective.**Minimize** to find the value(s) of (Gamma, C) that maximize the surrogate function of loss rate. For pos-pred % rate surrogate function, we use cvxpy.problems.objective.**Minimize** to find the value(s) of (Gamma, C) that minimize the surrogate function of pos-pred %. We can then add up these two sub-problems of minimization into one problem of minimization using the advanced features of CVXPY based on the documentation.

D. Performance Metrics

Since PU-learning is different from traditional machine learning like supervised learning with positive-negative data, we cannot use precision and recall rate to evaluate the performance of the Weighted Elkanoto Classifier. As a result, we proposed the following two metrics: loss rate and pos-pred %

- Loss rate is defined by the following equation:

loss rate = 1 - (number of predicted true positive data / number of true positive data in the testing set).

The rationale is that although we don't know if the predicted positive data from the unlabeled data is definitely true or not, the classifier should be able to correctly identify as many original positive labeled data set in the test set as possible. As a result, the lower the loss rate, the less true original positive labeled data is missing during the classification on the test set.

- pos-pred% is defined by the following equation:

pos-pred % = number of predicted positive data / number of total data in the testing set

The rationale is that the classifier is likely to predict all testing data to be positive so that it can obtain a 0% loss rate. However, this will make the classifier not robust at all. As a result, we want the classifier to predict as few positive data as possible from the unlabeled data as possible.

Then by combining the two metrics together, we want the classifier to use as few positive predicted data as possible to maximize its ability to reclaim all the original positive labeled data in the test set. As a result, the values of these two metrics adding up together will be the final value of measuring the performance of the Weighted Elkanoto Classifier with specific C and Gamma hyperparameters.

VI. Results

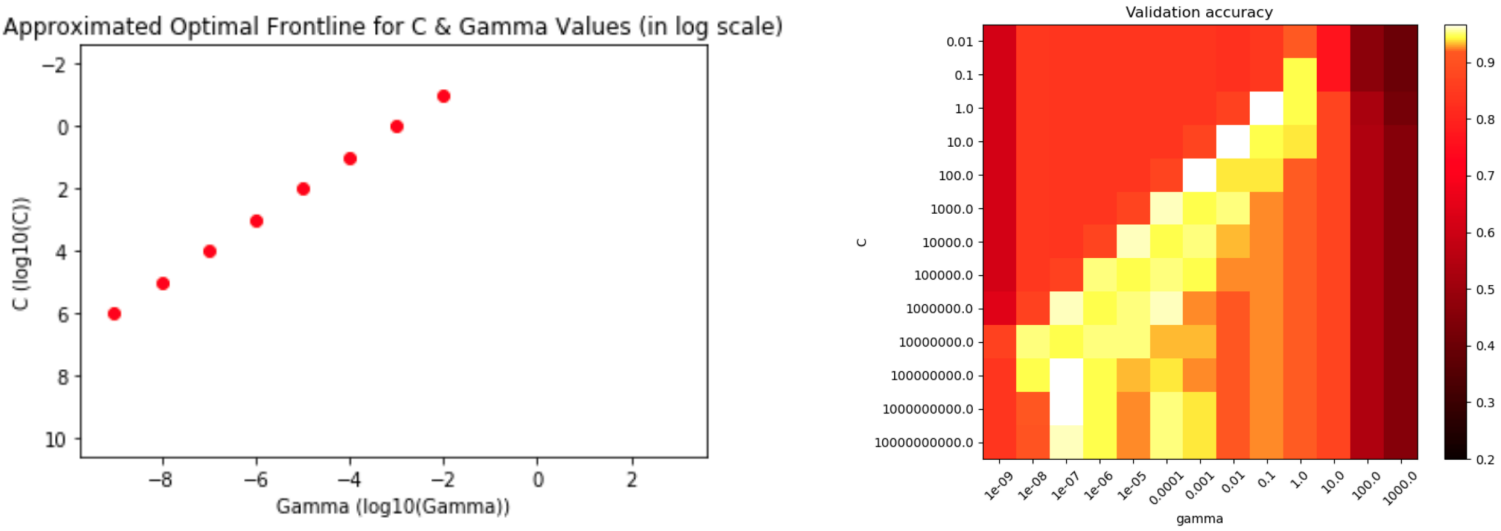

Fig. 2 Values of C-Gamma optimizing model performance (Left) and Original Predicted Accuracy on Scikit Learn (Right)

Noted in the right graph the lighter color means higher accuracy. And in the left graph, the red dots indicate the values of C-Gamma for which the Weighted Elkanoto Classifier performs the best in the As shown above, we can see that the C-Gamma hyperparameters inorganic material synthesizability prediction problem with 1:1 positive to unlabeled ratio.

The result is quite interesting since the optimal frontier we have for the Weighted Elkanoto Classifier is in the low accuracy region predicted by the Scikit Learn graph. Thus, the hyperparameter selections highly depend on the specific problem we are dealing with.

References

[1] Kochenderfer, M. J., & Wheeler, T. A. (2019). Algorithms for optimization. Mit Press.
[2] Elkan, C., & Noto, K. (2008, August). Learning classifiers from only positive and unlabeled data. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 213-220).
[3] SMT: M. A. Bouhlel and J. T. Hwang and N. Bartoli and R. Lafage and J. Morlier and J. R. R. A. Martins.
[4] https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html
[5] https://www.cvxpy.org/index.html