CHE597 - Final Paper

Active Learning

Sanjay Iyer

Active Learning

Active learning is a method that optimizes the way a machine learning model learns. Some data sets can be very large, which might make it really expensive to train. Other data sets might be very complex, making it hard to accurately train a model on its own. Active learning attempts to solve these two issues by having a human, which we call an annotator, occasionally check on the data while the model is learning.

The way this works is by first, starting with a small set of data labeled by an annotator. This should be a very small portion of the data, being around 1% of it. Next we train our model on this labeled data. Then we use this model to make predictions on the rest of our data. Any predictions the model is confident in, we keep. Any predictions the model is not confident in are sent back to the annotator to correctly label the data. Then we repeat the cycle and retrain the model using this newly labeled data. We continue through this loop until the model's predictions are robust. The level of robustness needed may differ depending on the circumstances.

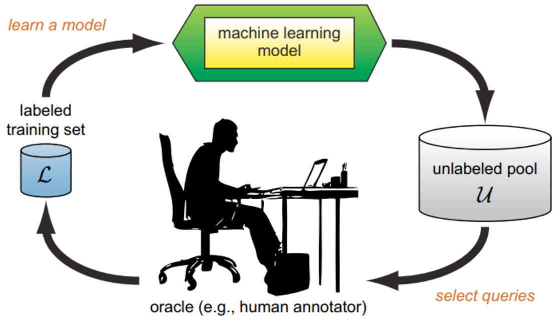

Figure 1 – Illustration of active learning loop.[1]

The activate learning model can use any known machine learning classifier. After selecting a classifier we can choose a framework, either being pooling or streaming. A streaming framework reviews samples one at a time, while a pooling framework reviews samples in batches.

Next we have to decide how we are going to have our model will decide if data is good or not. This can be done by setting a confidence interval, typically ninety-five percent. The other way is to do model by committee, which means it removes samples with the most disagreement across multiple models. The most commonly used combination of these is pooling with a confidence interval.

Active learning is a semi-supervised method that trains using labeled and unlabeled data. It is useful when working on a large unlabeled data set. Using active learning we can train a model with this data set faster and via less expensive means. Luckily, we don't have to start from scratch. If possible, using transfer learning to train our model initially would help expediate the training process.

One downside to active learning is it can only be used when our data is labeled. If we had no idea how to interpret the data, it would be very difficult to use active learning. Another problem is it gravitates towards outliers. If we randomly picked samples to label at the start and it had a few outliers, it may skew our data making erroneous data seem relevant. Currently active learning is not used very often. However, it appears to be gaining traction in the community.

Quantum Chemistry-Informed Active Learning to Accelerate the Design and Discovery of Sustainable Energy Storage Materials

Active learning was used in conjuncture with DFT simulations to compute the potentials of homobenzylic ether (HBE) molecules. They used Bayesian optimization for the model while trying to limit the number of DFT calculations. The goal in this was to discover novel molecules that can be used to design better energy storage materials. Whenever you are doing quantum calculations, it is likely that you will be doing expensive molecular simulations. Active learning was used in this research to help with this problem.

To train a model you need to select features. In this paper they selected physical descriptors including: molecular weight, topological surface area, number of valence electrons, and number of aromatic rings; for a total of 49 features. Then they reduced dimensionality to 15 principle components (PC) since the others were redundant and did not provide anything meaningful to the model.

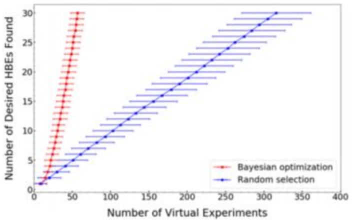

Figure 2 - Bayesian Optimization vs Random Selection[2]

As we can see from figure number 2, the Bayesian optimization model performed much better than random selection. This is evidence that the model improves selection and is not taking wild guesses. The importantance of this figure is to show that they were able to create a model using active learning that is faster than a standard machine learning model, while still being better than random.

It would have been helpful if they compared the Bayesian optimization to more than one other method. I believe most methods will perform better than random selection, infact I haven't seen any published paper showing a model performing worse than random selection.

One useful way to correlate the data was using a 2-D representation of the data with two different features for the x and y-axis. This is a helpful way to show a general overview of all the data points and see how all the compounds correlate to each other. In the paper they were most interested in compounds with the oxidation potential being between 1.4-1.7V for the HBEs. The model predicted nine percent of the HBEs would fall in this range. On top of this the model can discover patterns in the types of features that fall into their desired category. These patterns would be impossible to recognize without computational help. This will lead to faster discoveries in the field of material science.

Accelerated Discovery of Novel Inorganic Materials with Desired Properties Using Active Learning

Machine learning is being used to discover novel structures to be used in materials science. However current models are not always reliable and can have trouble making solid predictions that are fruitful experimentally. There are lots of large databases characterizing materials, but they vary in criteria and labels. This makes it difficult to throw all the data together and use one model without having any issues. The database is also too large to go through it with people. This is a perfect scenario where active learning shines. For this reason, they turned to active learning optimize a model in a faster and less expensive way.

Their reason for using active learning is that “by using descriptors to group the structures, with respect to their properties, a deeper understanding…in searching for target structures could be achieved.”[2] Finding structures with similar properties is key to developing novel materials. The bigger the bucket of leggos you give a chemist, the better the structure they will create.

They trained their data using roughly one to two percent of the database. Using the EGO method they were able to find the structure with the largest value in both criteria (band gap and refractive index), after searching only roughly seven percent of the database.

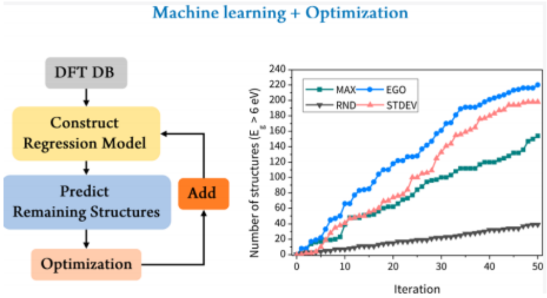

Figure 3 -Shows the looping system used and its outcome over each iteration using a variety of methods.[3]

This figure shows the number of structures found over each iteration that meet a specific criteria. This shows us that all optimization methods perform much better than random. It also shows us the MAX method is best for picking structures where the refractive index is greater than fifteen.

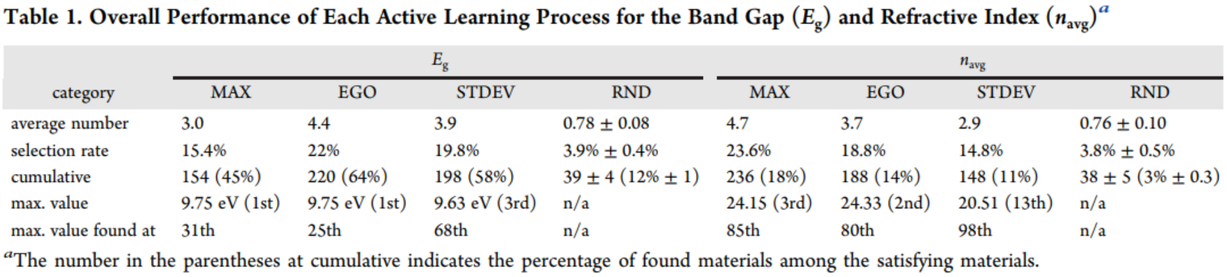

Table 1. Overall Performance of Each Active Learning Process for the Band Gap (E_g) and Refractive Index (n_{avg})[a]

	E_g				n_{avg}			
category	MAX	EGO	STDEV	RND	MAX	EGO	STDEV	RND
average number	3.0	4.4	3.9	0.78 ± 0.08	4.7	3.7	2.9	0.76 ± 0.10
selection rate	15.4%	22%	19.8%	3.9% ± 0.4%	23.6%	18.8%	14.8%	3.8% ± 0.5%
cumulative	154 (45%)	220 (64%)	198 (58%)	39 ± 4 (12% ± 1)	236 (18%)	188 (14%)	148 (11%)	38 ± 5 (3% ± 0.3)
max. value	9.75 eV (1st)	9.75 eV (1st)	9.63 eV (3rd)	n/a	24.15 (3rd)	24.33 (2nd)	20.51 (13th)	n/a
max. value found at	31th	25th	68th	n/a	85th	80th	98th	n/a

[a] The number in the parentheses at cumulative indicates the percentage of found materials among the satisfying materials.

This table shows that the model can find the max value for the band gap after only 25-68 iterations depending on the method. It can also find the max value of the refractive index after only 80-98 iterations. However the max value is not always the largest bandgap or refractive index. They don't explain why they stopped at the 13th largest structure instead of continuing until it was in the top three like every other method. They should have kept the goal uniform and attempted to get top three for each category.

References

(1) Science, O. (2018, December 12). Active learning: Your model's new personal trainer

(2) Doan, H. A., Agarwal, G., Qian, H., Counihan, M. J., Rodríguez-López, J., Moore, J. S., & Assary, R. S. (2020). Quantum Chemistry-Informed active learning to accelerate the design and discovery of sustainable energy storage materials. *Chemistry of Materials, 32*(15), 6338-6346. doi:10.1021/acs.chemmater.0c00768

(3) Min, K., & Cho, E. (2020). Accelerated discovery of novel inorganic materials with desired properties using active learning. *The Journal of Physical Chemistry C, 124*(27), 14759-14767. doi:10.1021/acs.jpcc.0c00545