Machine Learning techniques for predicting molecular properties

Viviana Petrescu
I&C, EPFL

.

***Abstract*—The high computational cost of quantum chemistry calculations have prompted the use of less expensive machine learning methods for predicting molecular properties in chemical compound space. Finding good feature representations for molecules is hard, in part because of the graph-like structure geometry of the molecules that need to be represented as high dimensional vectors. Add what this paper is about**

***Index Terms*—thesis proposal, candidacy exam write-up, EDIC, EPFL**

I. Introduction

The discovery of new molecular materials in chemistry has the potential of solving many of the problems we face today. Having a system which predicts both accurately and at a small computational cost the properties of new materials is highly desirable and has applications ranging from novel drugs discovery, water purification to efficient materials design for high energy transmission and storage [1].

II. Background Work

This write-up serves two purposes. First, it forms the basis for your candidacy exam. As such you should summarize the three papers selected by your advisor and yourself, and analyze as well as discuss them critically. Second, the write-up is also your thesis proposal. Therefore, the last one or two pages should be dedicated to your own preliminary work. A road-map of how you plan to advance the state of the art in your chosen area should also be given. For further details please consult the document "PhD Candidacy Exam Overview." You can find the latest version at `http://phd.epfl.ch/page57746-en.html`. Describe briefly the context, the problem, shortcomings in prior approaches, and your proposed approach and solution. Forecast results. Background — Describe the three papers in detail, the problem they tackle, the solutions and results, and their shortcomings, and how they relate to your work. This part builds the basis for the oral candidacy exam.

A. Learning Invariant Representations of Molecules for Atomization Energy prediction

Representing Molecules

While domain specific descriptors for molecules exist [4], recent work [3] has proposed to predict properties of a molecule of size N only from the 3D positions of the atoms R_i $i \in 1..N$ and their nuclear charge Z_i $i \in 1..N$. This has prompted the introduction of the Coloumb Matrix descriptor, whose individual entries appear in the Schroedinger equation. It is defined by a NxN matrix with entries given by:

$$M(i,j) = \begin{cases} 0.5 * Z_i^{2.4} & if i = j \\ Z_i * Z_j / |R_i - R_j| & otherwise \end{cases} \quad (1)$$

where $Ri - Rj$ represents the distance between the atoms i and j.

The dimensionality of the Coloumb Matrix is given by the number of atoms in a molecule. Since that varies across molecules in a dataset, one common trick is to pad with 0's the matrices corresponding to small molecules until they reach the maximum molecule size in a dataset. This limits the size of the molecules that can be used, since the descriptor has complexity $O(N^2)$, where N is the number of atoms.

Desired properties of descriptor Due to the graph-like structure of the molecules, finding a fixed size representation is difficult. The desired properties of a molecular descriptor are invariance to translation and rotation of the molecule and invariance to the indexing of the atoms.

Solved using sorted or Random Coloumb While the Coloumb Matrix representations solves the rotation and translation invariance through the use of distances between atoms

Proposal submitted to committee: June 13th, 2009; Candidacy exam date: June 20th, 2009; Candidacy exam committee: Exam president, thesis director, co-examiner.

This research plan has been approved:

Date: _____

Doctoral candidate: _____
(name and signature)

Thesis director: _____
(name and signature)

Thesis co-director: _____
(if applicable) (name and signature)

Doct. prog. director: _____
(B. Falsafi) (signature)

$R_i - R_j$, invariance in atom indexing still needs to be tackled since any permutation of atom indexes results in a valid Coloumb descriptor. Two variations of the descriptor are proposed in [2]. The first one, called Sorted Coloumb, uses the permutation of the atoms given by the sorting of the row norms of a valid Coloumb matrix. Any molecule has a unique representation given by the Sorted Coloumb matrix. The second representations is based on the idea that the norms of the rows can have very similar values in practice and the sorting, therefor also the atom indexing, is subject to small noise. The new descriptor, Random Coloumb Matrices, extends the Sorted Coloumb matrix and samples multiple Sorted Coloumb matrices, were the set of rows was sorted according to their norm at which a small Gaussian noise was added.

Reach state of the art result at the time. The current new value is 1.5 using Bag of Bonds.

B. Self-Taught Learning: Transfer Learning from Unlabeled Data

Present the different machine learning framewroks that exists, with picture. It is a two part optimization problem. - First find bases - Then find new activations for the old ones. Through experimental results - since they were tested on a range of problems from image, text and sound.

C. Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting

Optimizing an unknown function is an important problem. Often done using Gaussian processes in a Bayesian setting. Many heuristic have been proposed for defining an acquisition function. Few is known for its conergence. Sublinear converge rates are proven for GP-UCP. The bound is in two steps. First bound it on the information gain, then on the spectral of the kernel matrix which is further bound on the spectral operator.

III. Research Proposal

A.

Proposing a new descriptor which is invariant to permutations of the atom. 2.6+/ or 2.1 with noise. (still gives comparable accuracy to same DFT models)

B.

Propose of augumententing the data sets more easily and make the cross validation less sensitive to splitting -at the moment : take non H atoms, then sort then do CV.

Pose the problem as one of the semi supervised, self tought learning etc problems and try to make it generalize across compound space, or learn new embeddings of the atoms.

Although the use of Gaussian processes in material design is not new, it s major drawback is the computational bottleneck. In our scenario, this can be used for tuning the hyper-parameters of the model trained. Here the input dimensionality is given by the nbr of the hyper parameters (learning rate, activation fct, nbr hidden layers) used and the fct to be minimized is the cross validation error. The result presented in the previous subsection were obtained like that. Talk if we have time about bayesian neural nets, were simpler models outperform more easy models just be using hyper parameter optimization instead of grid search.

Write how you propose to advance the state of the art given the background. What is new technically? How does it improve over prior work? Summarize, suggest an approximate timeline, and list references.

References

[1] Johannes Hachmann, Roberto Olivares-Amaya, Sule Atahan-Evrenk, Carlos Amador-Bedolla, Roel S. Snchez-Carrera, Aryeh Gold-Parker, Leslie Vogt, Anna M. Brockway, and Aln Aspuru-Guzik. The harvard clean energy project: Large-scale computational screening and design of organic photovoltaics on the world community grid. *The Journal of Physical Chemistry Letters*, 2(17):2241–2251, 2011.

[2] Grégoire Montavon, Katja Hansen, Siamac Fazli, Matthias Rupp, Franziska Biegler, Andreas Ziehe, Alexandre Tkatchenko, Anatole V Lilienfeld, and Klaus-Robert Müller. Learning invariant representations of molecules for atomization energy prediction. In *Advances in Neural Information Processing Systems*, pages 440–448, 2012.

[3] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.*, 108:058301, Jan 2012.

[4] Roberto Todeschini and Viviana Consonni. *Handbook of molecular descriptors*, volume 11. Wiley-VCH, 2000.