Deep learning classification tool for x-ray scattering data of cellulose molecules

Yan Zhang
Electrical and Computer Engineering, Northeastern University Boston, MA
yzhang@ece.neu.edu

12/06/2016

***Abstract*—Two major type of machine learning models were developed to classify the x-ray scattering intensity data of cellulose molecules: support vector machine based method and neural network based method.**

***Keywords*—Deep support vector machine; Convolutional neural network**

I. Introduction

This is an interdisciplinary study includes three major research areas:

- *X-ray scattering in Photonic Physics*
- *Molecular dynamics in Biochemistry*
- *Machine Learning in Computer Science*

X-ray scattering technique has been used to investigate molecular structure in micro/nano length-scale in biomass-biofuel transformation. The diffraction patterns collected at synchrotron source indicate the averaged shape and size of cellulose molecule in maize stover. In order to further understand the details of cellulose fibrous molecules, a large number of molecular structures of cellulose were simulated using molecular dynamic program Charmm. This molecular structure dataset includes 11 shapes (labels) and 2000 structural variation for each shape. Scattering intensities and diffraction patterns of each structure were simulated using Debye formula which is basically a fourier transform of the autocorrelation function of electron density of a molecule. The molecular structure characterization problem is now being transformed into a machine learning and pattern recognition problem - training a classifier using simulated intensities and labels, then classifing the experimental intensity.

II. Deep support vector machine

First method is Deep-SVM, a hierarchical decision-tree based multi-class support vector machine model. This model first divides the whole dataset into two categories and then continues to perform binary classification for each layer categories until the bottom child layer was reached. If the number of labels is N, then $N-1$ classifiers needs to be trained. This method has advantages over traditional two type of SVM classifiers: one vs. all and one vs. one method. One vs. all method has huge bias when the number of labels is increasing and the number of classifiers need to be trained is increasing quadratically for one vs. one method ($\frac{N(N-1)}{2}$). The Deep-SVM model can reduce both system bias and number of classifiers.

III. Multilayer perceptron and convolutional neural network

Second method is generalized neural network framework which can be constructed from shallow to deep learning architecture:

- *Softmax Regression*
- *Multi-layer Perceptron*
- *Convolutional Neural Network*

If the classification system only has input and output layers plus activation function, it is softmax regression model. The multi-layer perceptron model can be built by adding hidden layers, which has better performance on large dataset and higher dimensional data. Experimental intensity data (test data) often has different type of noise and intense background scattering intensity. Several convolutional and pooling layers are added in order to enhance the signal-to-noise ratio and extract features of molecular scattered intensity, which makes the system become a convolutional neural network (deep learning) model.

Acknowledgment

This work was supported as part of the Center for Direct Catalytic Conversion of Biomass to Biofuels (C3Bio), an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Basic Energy Science under award #DE-SC0000997.

References

[1] Pedregosa, Fabian, et al. ”Scikit-learn: Machine learning in Python.” Journal of Machine Learning Research 12.Oct (2011): 2825-2830.

[2] Buitinck, Lars, et al. ”API design for machine learning software: experiences from the scikit-learn project.” arXiv preprint arXiv:1309.0238 (2013).

[3] Jia, Yangqing, et al. ”Caffe: Convolutional architecture for fast feature embedding.” Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014.

[4] Team, The Theano Development, et al. ”Theano: A Python framework for fast computation of mathematical expressions.” arXiv preprint arXiv:1605.02688 (2016).

[5] Abadi, Martn, et al. ”Tensorflow: Large-scale machine learning on heterogeneous distributed systems.” arXiv preprint arXiv:1603.04467 (2016).

[6] Collobert, Ronan, Koray Kavukcuoglu, and Clment Farabet. ”Torch7: A matlab-like environment for machine learning.” BigLearn, NIPS Workshop. No. EPFL-CONF-192376. 2011.

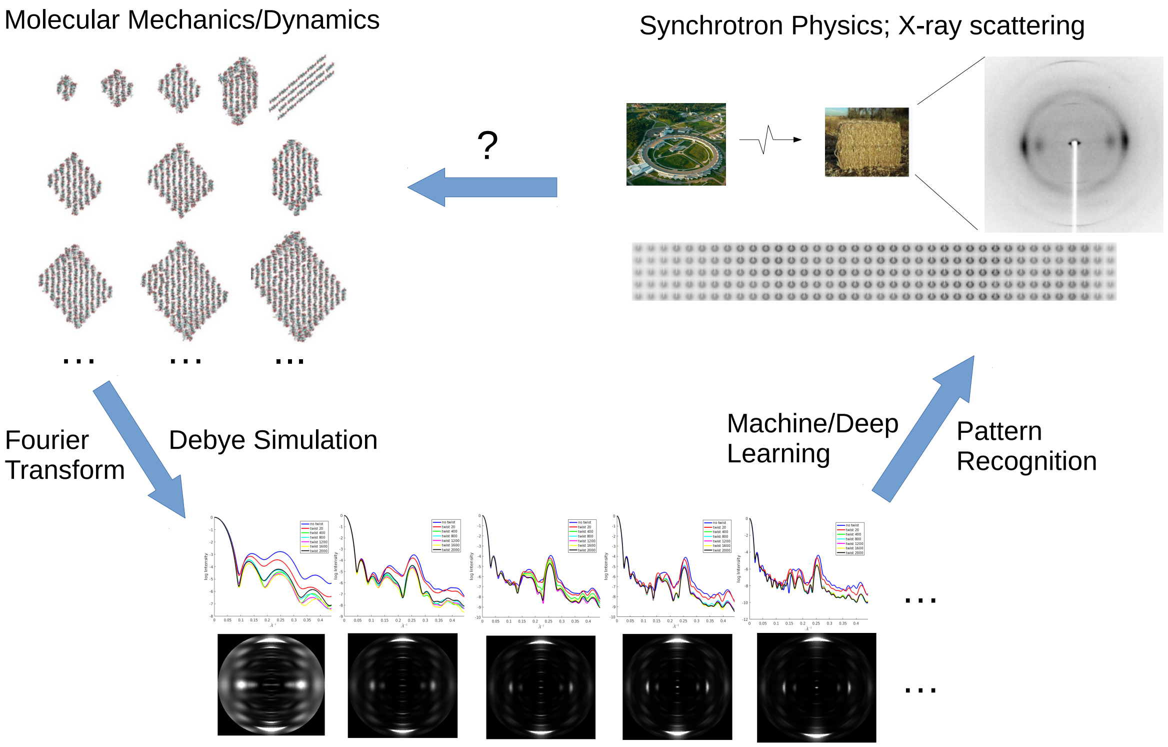

Fig. 1: Outline of classification system. Top-left: Cellulose molecular structures simulated by Charmm molecular mechanic progrem. Top-right: Experimentall diffraction patterns of milled maize stover collected at Syntrochron. Bottom: Simulated intensities and diffraction patterns using simulated molecular structures.

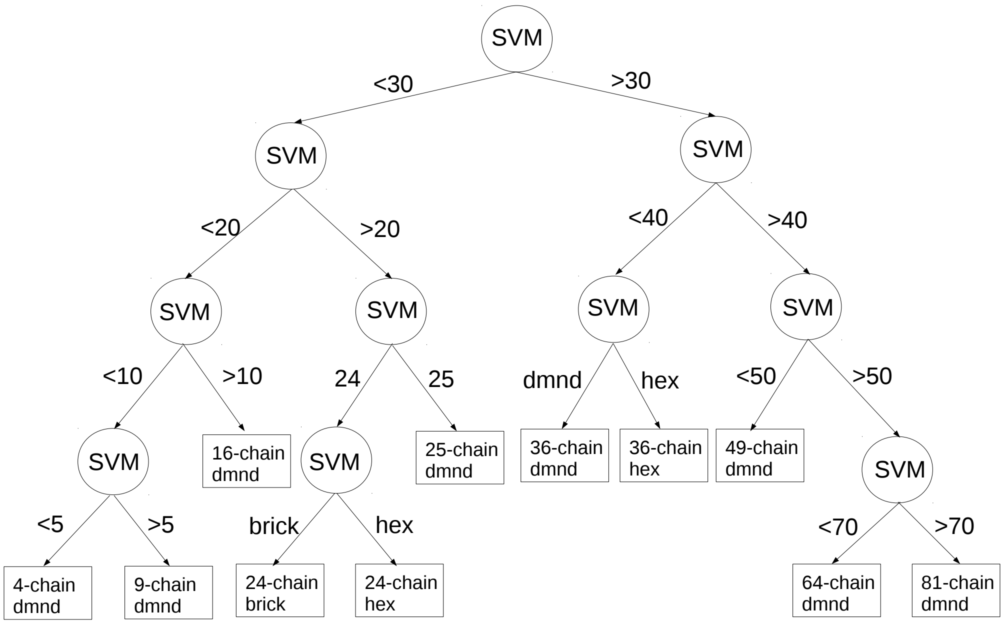

Fig. 2: Deep-SVM: A hierarchical decision-tree based multi-class support vector machine model.

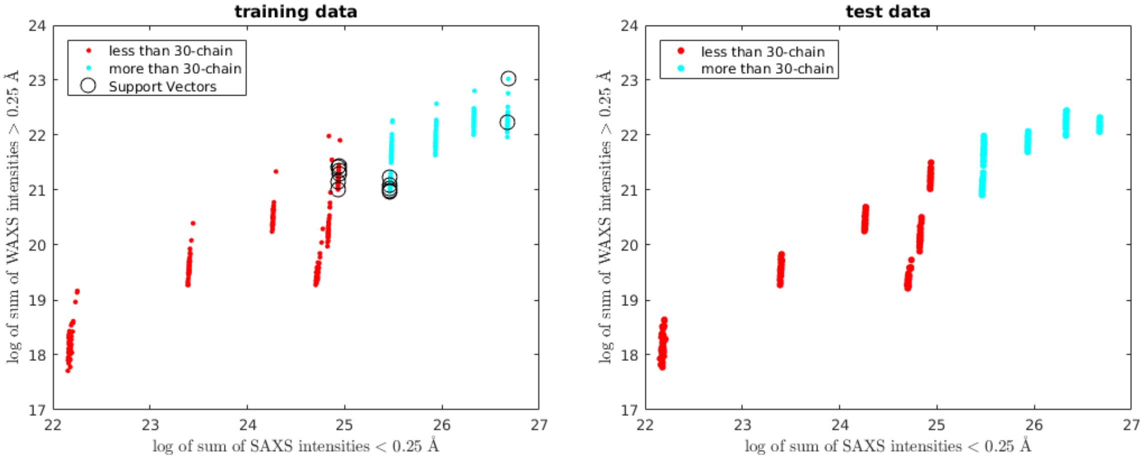

Fig. 3: Classification using Deep-SVM. Left: First layer SVM training result. Right: First layer SVM test result.

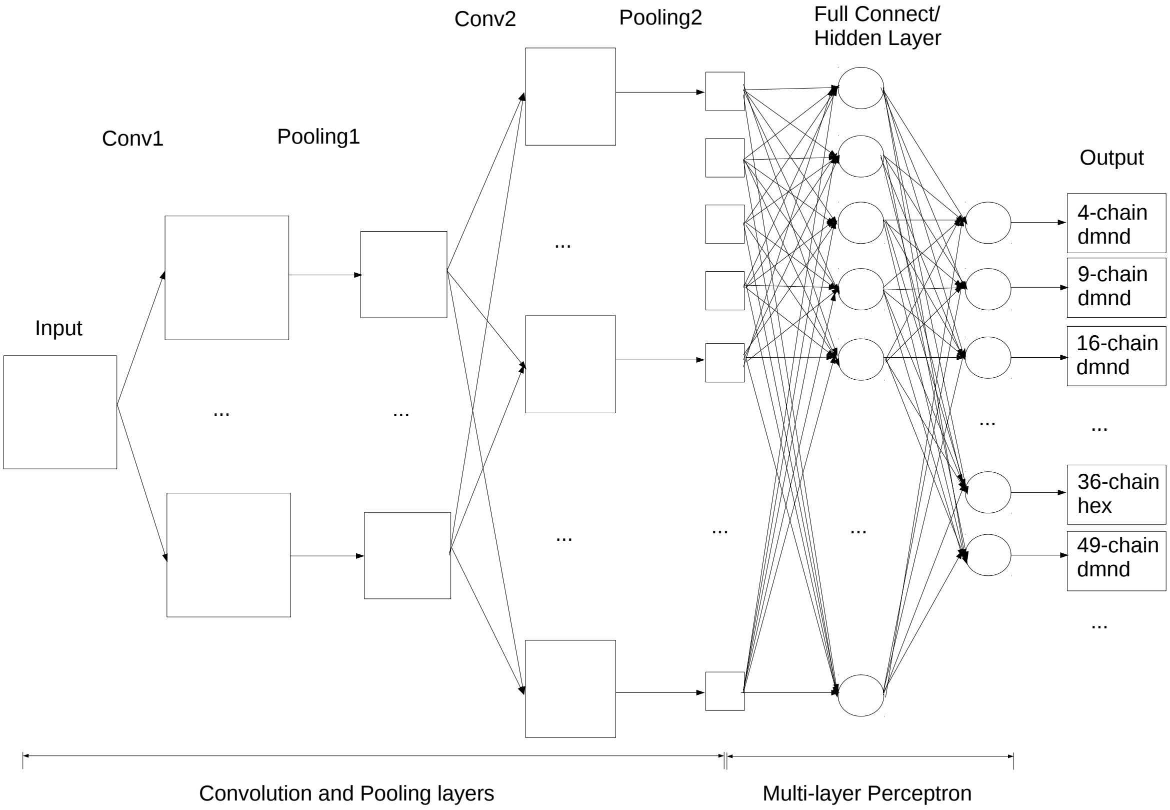

Fig. 4: Multi-layer perceptron and convolutional neural network framework.