Multi-Contextual Representation and Learning with Applications in Materials Knowledge Discovery

Ph.D. Dissertation Proposal

Ruoqian Liu

EECS, Northwestern University

rosanne@northwestern.edu

May 19, 2015

Overview

Outline

Machine Learning towards Multiple Contexts

- Data mining is largely an empirical science.
- The quality of data plays an important role in the performance.
- Data has become so big and complex.
- Data has a **multi-contextual** nature.

Machine Learning towards Multiple Contexts

- Data has a **multi-contextual** nature.

Machine Learning towards Multiple Contexts

- Data has a **multi-contextual** nature.

House price prediction

Local predictors:

crime rate, pollution, number of rooms, school district, property tax rate, …

Higher-level factors:

job market stability, political stance, city team Super Bowl appearance, …

Different Contexts!

Reviews sentiment analysis

amazon

Kitchen appliances:

“malfunction”, “reliable”, “sturdy”, …

DVDs:

“thrilling”, “horrific”, “hilarious”, …

Different Contexts!

Machine Learning towards Multiple Contexts

Definition of a **"Context"**:

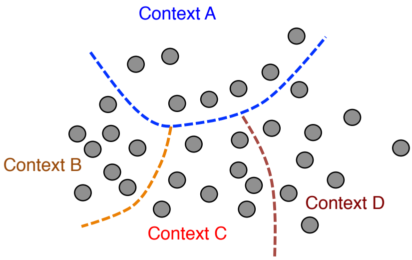
- A context is a high level learning environment.
- Data distribution are different across contexts.
- Multiple contexts are not clearly annotated.

Definition of a **"Multi-contextual problem"**:

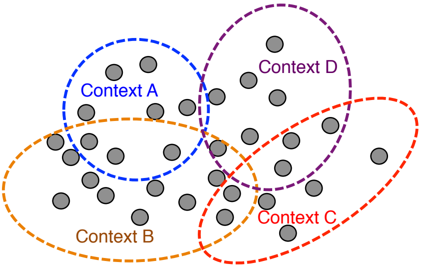
- The learning problem has data collected from multiple contexts.
- The test data presents a distribution largely different from training,
- or, the test data contains multiple distributions as well.
- Prior knowledge of such distributions may be unavailable.

Multi-Contextual in Data Level

The multi-contextual quality may exist in data.



(a) Multi-contextual data: to partition

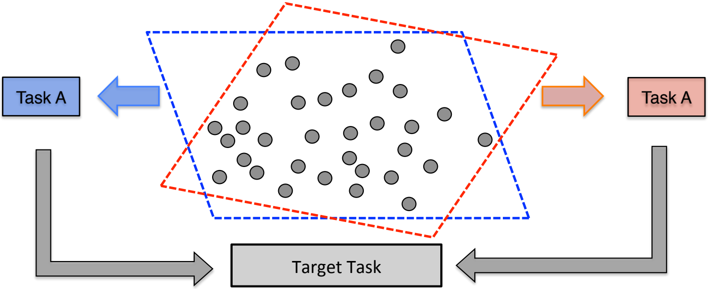


(b) Multi-contextual data: to resample

An intuitive idea: partition the data, into non-overlapping subgroups; or resample the data.

Multi-Contextual in Task Level

The multi-contextual quality can also exist in the task level.



This is particularly useful in high-dimensional datasets, where some dimensions can be abandoned in formulating sub-tasks.

Research Objective

The objective of this dissertation is to develop advanced data mining techniques, to manage the information complexity lying in scientific data such as materials structural data; identify, extract and address the multi-contextual characteristics in materials knowledge discovery problems.

Those problems often present themselves as a single task. However, because of the data structural complexity that is representative in materials systems, a single, flat, shallow structured model is insufficient in providing an accurate approximation. We therefore claim that they all bear a multi-contextual quality, which should be addressed in a hierarchical, multi-layered fashion.

The contextual difference between test data and training data is to be addressed by unsupervised learning.

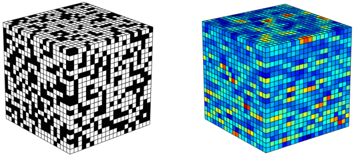
Advantages of Building Models for Each Context

- During training, each model need only consider training examples of its own context, saving an order of magnitude in computation, with no reduction in discriminative performance.
- After training, it is possible to add a new context without retraining the previous context models.
- It is possible to fit far more parameters before overfitting occurs because the input vectors contain much more information than the class label.
- Single density models can be fitted by methods like expectation-maximization (EM) that are considerably more efficient than gradient descent.

Case Study

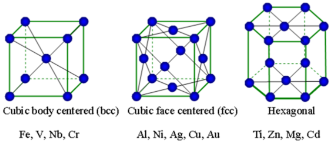
Structure-property modeling of binary composites

Efficient prediction of heterogeneous elastic fields in a three dimensional (3-D) voxel-based microstructure volume element.



Quantum Crystalline Compounds Property Prediction

Looking at the composition and structure of crystalline compounds; making predictions about their formation energy.



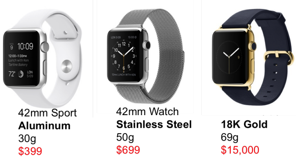
Outline

What is Materials Science and Engineering?

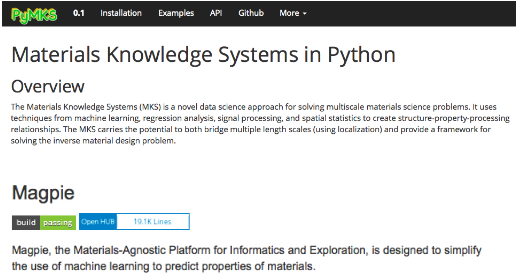
We CS don't know MSE well...

Copper age? Bronze age? Iron age?

Silicon valley?



But they know us...


PyMKS 0.1 Installation Examples API Github More

Materials Knowledge Systems in Python

Overview

The Materials Knowledge Systems (MKS) is a novel data science approach for solving multiscale materials science problems. It uses techniques from machine learning, regression analysis, signal processing, and spatial statistics to create structure-property-processing relationships. The MKS carries the potential to both bridge multiple length scales (using localization) and provide a framework for solving the inverse material design problem.

Magpie

build passing Open HUB 19.1K Lines

Magpie, the Materials-Agnostic Platform for Informatics and Exploration, is designed to simplify the use of machine learning to predict properties of materials.

Materials Informatics: Data Mining in Materials Science

Material science is about:

- How materials are put together
- How they can be changed to improve
- How to discover new

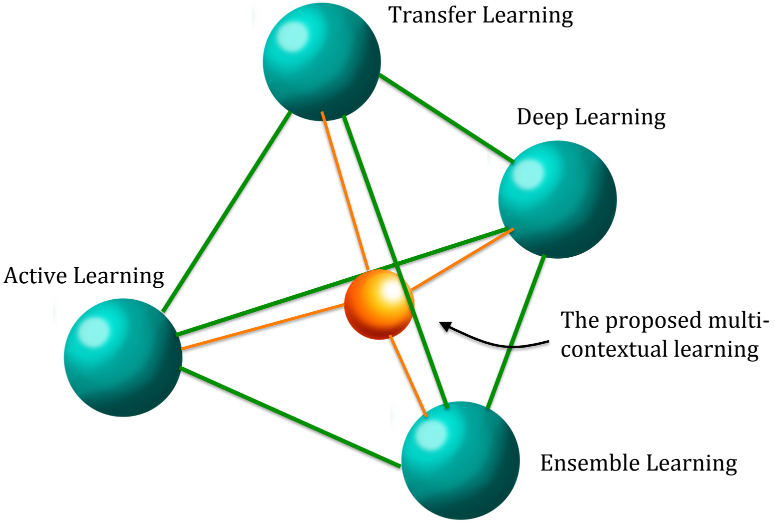
Before informatics, the answers are majorly driven by trail-and-error.

A large number of trails only give a small portion of infrequent and unexpected discoveries.

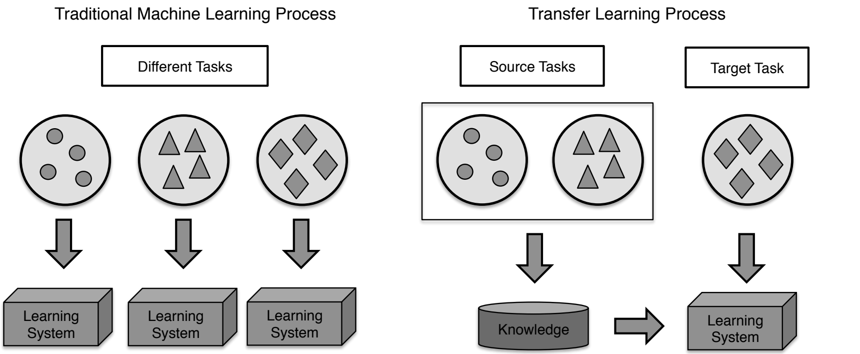
The goal of **data mining** in materials informatics:

- Integrating materials information across multiple length scales.
- Providing an accelerated means of recognizing structure-property relationships.

Related Fields of Machine Learning in Proposed Work



Transfer Learning



Reproduced from: Pan, S. J., & Yang, Q. (2010). "A survey on transfer learning". *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345-1359.

Deep Learning

Learning the representation

- We know the way in which data are represented can make a huge difference in the success of a learning algorithm.
- Deep learning enables the learning of multiple levels of representation, discovering more abstract features in the higher levels.

Learning as human does

- Because human brains appear deep, AI-tasks require deep circuits
- Because it is natural for humans to represent concepts at multiple levels of abstractions, deep architecture makes sense.
- Because human learn mostly unsupervised, only partially supervised.

Deep Learning: the Basic Recipe

Greedy Layer-Wise Learning of Representations

1. Let $h_0(x) = x$ be the lowest-level representation of the data, given by the observed raw input x.
2. For $l = 1$ to L

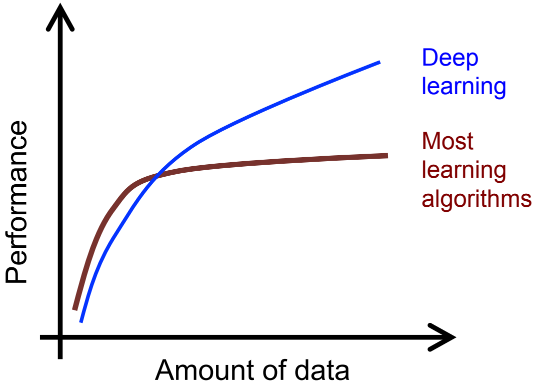
   Train an unsupervised learning model taking $h_{l-1}(x)$ at level $l-1$ as input, and after training, producing representations $h_l(x) = R_l(h_{l-1}(x))$ at the next level.

Several variants from this point on.

Supervised learning with fine-tuning: most common

Unsupervised: Deep autoencoders or a Deep Boltzmann Machine

Deep Learning: the Power of Handling Big Data



Reproduced from Andrew Ng's Invited Talk at RSS2014

Outline

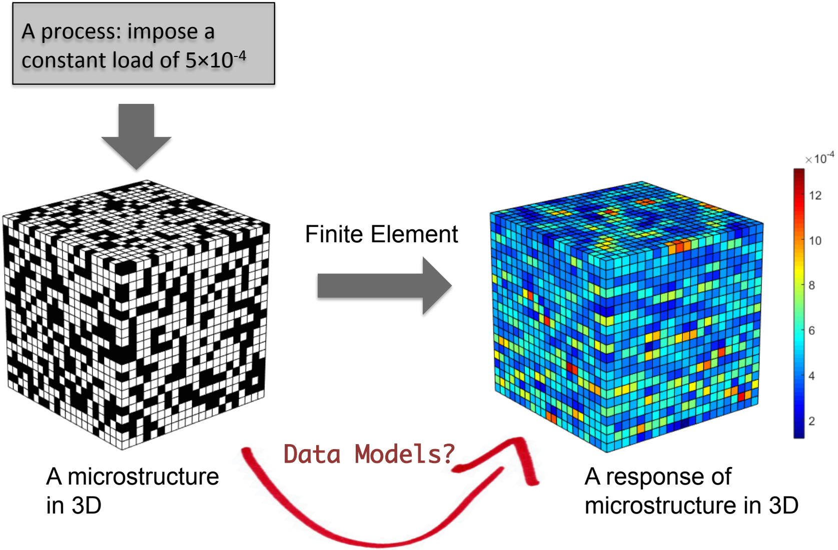
Structure-Property Modeling of Binary Composites

Problem

- Studies the **localization** effect and linkage towards elastic deformation of 2-phase composites.
- **Localization**: the spatial distribution of the response at the microscale for an imposed loading condition at the macroscale.
- Simulated output from physics-based materials finite element (FE) models are available, however, the computational requirements are very high.

Data

- A set containing digitally created 3D microstructure volume elements (MVEs) with each microstructure associated with a response field.
- Each MVE consists of $21 \times 21 \times 21$ binary-phase elements.
- The response field is captured as a continuous number in each spatial voxel.
- A total of 2500 MVEs with varying volume fractions are included.


A process: impose a constant load of 5×10-4
Finite Element
Data Models?
A microstructure in 3D
A response of microstructure in 3D
×10-4
12
10
8
6
4
2

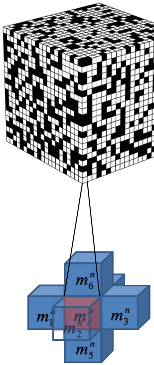
Method

Data Representation

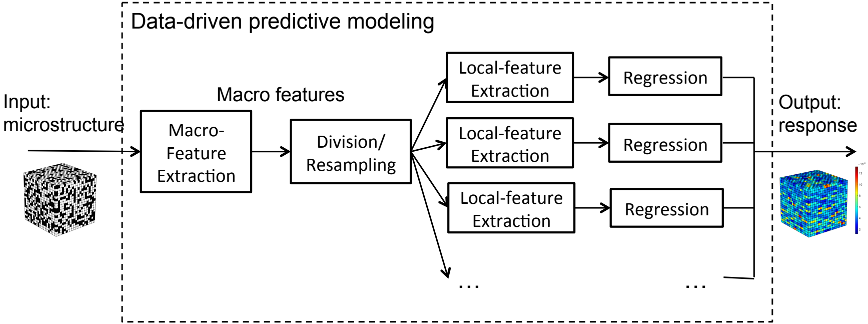
- Each voxel must be represented, by its neighboring information.
- What about the cube-wise information, like volume fraction?

Consider the “type of cube” as the context that voxels are drawn from.

- Use macro-features that represent the context.
- Build prediction model for each context.



Method



Macro-feature extraction: cube-wise characteristics, e.g., volume fraction, number of connected-same-phase components, ...

Division/resampling: Data division/resampling based on macro-features.

Local-feature extraction: voxel-wise characteristics. e.g., neighbor voxels.

Preliminary Results

Multiple contexts addressed by data partition:

P1: Partition based on volume fraction only (100 groups by default)

P2: Partition based on clustering of a set of macro-features (93 groups)

P3: Partition based on PCA of 2-point correlation functions (90 groups)

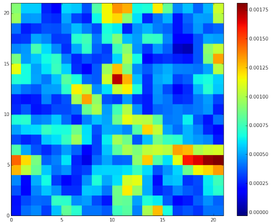
Contrast 10 System	**MASE (%)**
VF (100 groups)	8.89
Macro-features (93 groups)	8.43
PCA of 2-point correlation function	8.03

$$MASE = \frac{1}{S}\sum_{s=1}^{S}\left|\frac{p_s - \hat{p}_s}{p_{imposed}}\right| \times 100\%$$

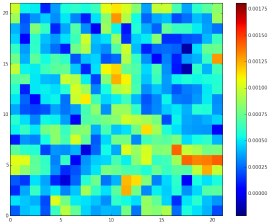
PCA of 2-point Correlation Function

A k-means like algorithm for partitioning data cubes into sub-model groups.

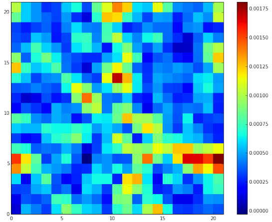
1. Decide on k sub-models;
2. Choose initial assignments among the k sub-models for each data cube example in the training set, typically at random;
3. Perform PCA separately for each sub-model;
4. Reassign data cubes to the sub-model that reconstructs them the best;
5. Stop if no examples have changed sub-model, otherwise return to 3.



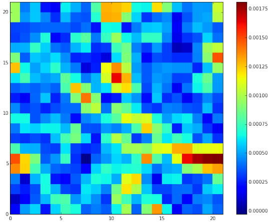
FEM
VF: 50.22%



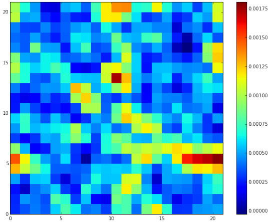
One Context, Single Model
All test error: 17.01
Cube error: 25.86
Slice error: 27.02



Multi-Context, P1
All test error: 8.89
Cube error: 12.32
Slice error: 12.13



Multi-Context, P2
All test error: 8.43
Cube error: 12.66
Slice error: 12.69



Multi-Context, P3
All test error: 8.03
Cube error: 12.14
Slice error: 11.93

Quantum Crystalline Compounds Property Prediction

Problem

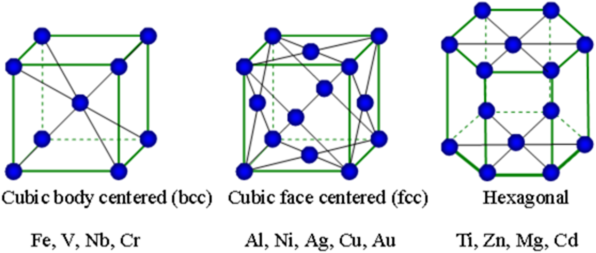
- Make use of exisiting database to identify common trends in material properties, and using those trends to create design rules.
- Predictors include both composition-based and structure-based attributes.
- Predicts the formation energy, the energy stored (presented by a negative, as opposed to released) when a compound is produced.
- Once predictable, we can predict the stability of new compounds.

Data

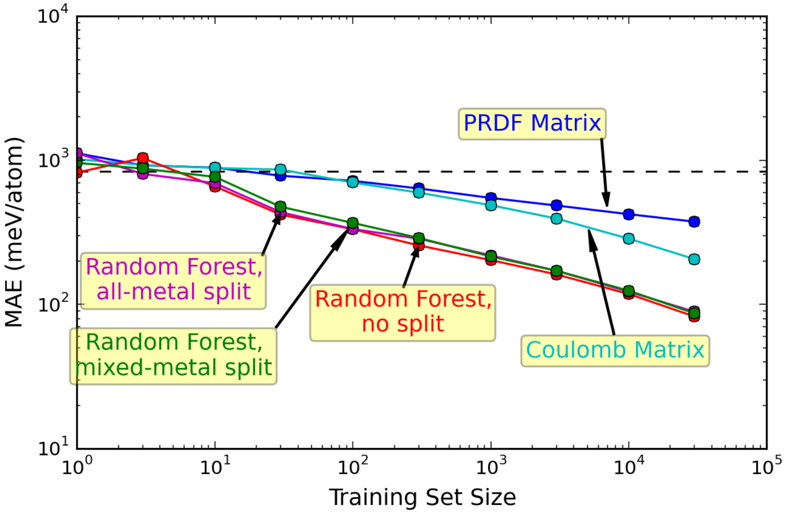
- The Open Quantum Materials Database (OQMD) contains the properties of around 300,000 crystalline compounds calculated using Density Functional Theory (DFT).
- Around 250 attributes are generated based on Voronoi tessellation.
- The formation energy is presented as a numerical value from -5 to 5 eV/atom.

The Underlying Multi-Contextual Problem

- All the compounds the world knows... how many contexts are there?
- Possible contexts, based on composition:
  - Does the compound contain metal elements?
  - Does the compound contains a halogen, chalcogen, or pnictogen?
  - Does the compound have a relatively small band gap energy?
  - ...
- Possible contexts, based on structure:
  - Which standard structural group does the compound belong to?
  - ...



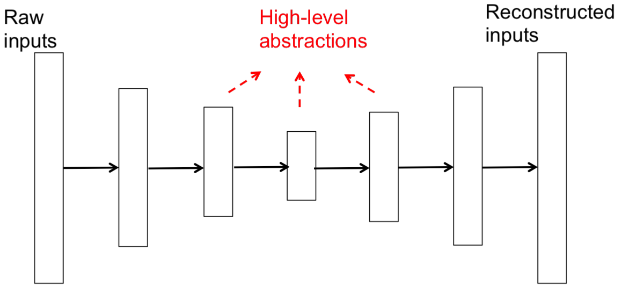
Preliminary Results on Simple Partitioning



Outline

Representation Learning Based Context Detection

- Use unlabeled data to extract high-level features;
- Build intermediate abstractions that are **shared** and meaningful across contexts.
- Detect context change with these abstractions.



Representation Learning: A Stacked Autoencoder

The Proposed Idea

Representation learning with data across multiple contexts

Representation-learning would discover features that capture the generic factors of variation present in all the classes.

Use the test set to select useful representations

Out of a large set of general-purpose features that covered the variations across many contexts, we select those few factors varying most in the test set (unlabeled).

The Proposed Idea (Continued)

Examine learned weights to disentangle different contexts

Partition/resample the training data to have multiple sub-models pre-trained with deep learning. The weights fixed by each should reveal the contextual information.

Examine reconstruction errors to assign test examples to sub-models

Test examples can be used in an unsupervised fashion. By applying sub-models learned from training data for reconstruction, the errors will tell whether the related context is appropriate.

The Links to Other Subfields of ML

Links to Active Learning

How to actively select data subsets that form a proper context.

Links to Ensemble Learning

How to aggregate multiple sub-models from each context.

Links to Transfer Learning

How to create source and target tasks.

Future Work Plan

Structure-Property Prediction for 3D Composites

- In identifying multiple structural-based contexts with deep networks, the 3D structure has to be taken into consideration.
- The challenge of massive data volume is still profound.

Compound Energy Prediction

- Partition compounds purely based on data statistics and verify with theory-based partition.
- Each compound group takes a subset of features that vary the most.

Financial Risk Assessment through WSJ News Title Analysis

- Identify multiple contexts as market environments.
- Associate environmental indicator with risks embedded in stock/futures/indices price and volume volatility.

More to explore: Does generating intermediate **tasks** help?

Almost The End

A Note to Myself

The past couple of years have been multi-contextual for me.

> We have one sole goal: to finish the Ph.D well.
>
> But we get so much data that come from so many contexts: various research projects, classes, teaching, attending talks, giving talks, social networking, . . .

With these multi-contextual data,

> *How to better utilize them?*
> *How to form different tasks with them?*
> *How to transfer knowledge learned from one task to another?*
> *How to combine all the knowledge acquired towards the final goal?*

As data mining, the answer is in the doing, maybe.

The End

The End

Thank You!