Bioinformatics Center
– Bio-knowledge Engineering –

https://www.bic.kyoto-u.ac.jp/pathway/index.html

Prof
MAMITSUKA, Hiroshi
(D Sc)

Senior Lect
NGUYEN, Hao Canh
(D Knowledge Science)

Student

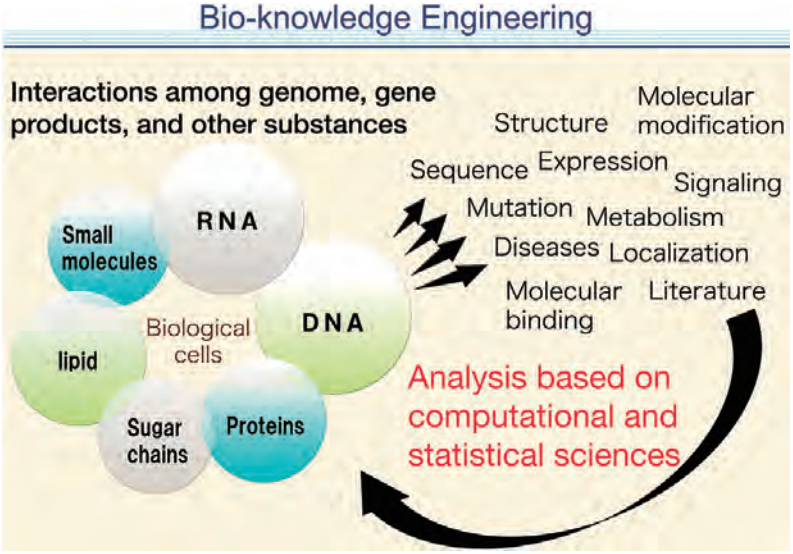
NGUYEN, Duc Anh (D3)

Guest Res Assoc

PETSCHNER, Peter (Ph D) Semmelweis University, Hungary, 28 November 2020– 27 November 2022

Scope of Research

We are interested in graphs and networks in biology, chemistry, and medical sciences, including metabolic networks, protein-protein interactions and chemical compounds. We have developed original techniques in machine learning and data mining for analyzing these graphs and networks, occasionally combining with table-format datasets, such as gene expression and chemical properties. We have applied the techniques developed to real data to demonstrate the performance of the methods and find new scientific insights.



KEYWORDS

Bioinformatics
Computational Genomics
Data Mining
Machine Learning
Systems Biology

Selected Publications

Wimalawarne, K.; Yamada, M.; Mamitsuka, H., Scaled Coupled Norms and Coupled Higher-Order Tensor Completion, *Neural Computation*, **32,** 447-484 (2020).

Nguyen, C. H., Structured Learning in Biological Domain, *Journal of Systems Science and Systems Engineering*, **29,** 440-453 (2020).

Nakamura, A.; Takigawa, I.; Mamitsuka, H., Efficiently Enumerating Substrings with Statistically Significant Frequencies of Locally Optimal Occurrences in Gigantic String, *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2020)*, **34(4),** 5240-5247 (2020).

Nguyen, C. H.; Mamitsuka, H., Learning on Hypergraphs with Sparsity, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (in press).

Wimalawarne, K.; Mamitsuka, H., Reshaped Tensor Nuclear Norms for Higher Order Tensor Completion, *Machine Learning* (in press).

Learning on Hypergraphs

Relational data has been of much interest in statistics and machine learning. It is the data with relationships among objects. To study this relational data, graph and hypergraph have been the main tools. Graph encodes relationships of pairs of objects. Hypergraph is a more general way of encoding high-order relationships, that is the relationship of variable numbers (two or more) of objects. It is a generalization of graph, in which only pairwise relationships can be represented. It finds applications in various domains where relationships of more than two objects are observed. In social networks, it can represent many groups of individuals that have common interests. In computer vision, it can represent patches of neighbor pixels that have similar colors. In bioinformatics, hypergraphs can be used to represent relationships among proteins in protein complexes, among sets of drugs and their side effects. In the example in Figure 1, we show a hypergraph that houses in the same street belongs to a set call a hyperedge. This is useful to study problem such as predicting house prices given various information including streets that the houses are located in.

On a hypergraph, as a generalization of graph, one wishes to learn a smooth function with respect to its topology. This is the semantics of graphs and hypergraphs in many domains. In social networks, individuals in the same academic groups might have interest in the same books. Patches of pixels with the same color in an image tends to belong to the same object in the image. Proteins in the same complex are likely to participate in the same cellular functionality. These observations have been formulated in statistics and machine learning, that labels of objects in the same given set (hyperedge) tend to be similar. Such a function is called *smooth on the graph/hypergraph*. It is a fundamental issue is to find suitable smoothness measures of functions on the nodes of a graph/hypergraph to make the function smooth.

There are various methods proposed to model hypergraphs. However, it is not clear why a method works for one problem, not for another problem. Strengths and weaknesses of the methods are hardly studied and poorly understood. In order to clarify all these methods, we show a general framework that generalizes all previously proposed smoothness measures on hypergraphs. Our framework not only allows for analyzing previous smoothness measures, but also gives rise to many new measures with useful properties. For example, we found that the commonly used total variation method to be too sensitive noises in data, which is very common in practice.

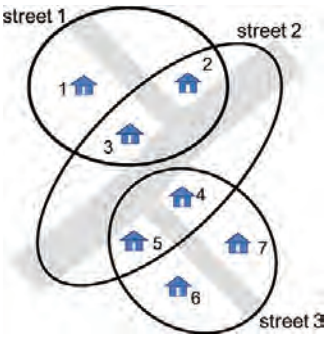


Figure 1. Houses on streets as nodes on hyperedges of a hypergraph.

To address the problem of irrelevant or noisy data, we wish to incorporate sparse learning framework into learning on hypergraphs. From our proposed framework, we propose sparsely smooth formulations that learn smooth functions and induce sparsity on hypergraphs at both hyperedge and node levels. We show their properties and sparse support recovery results. This is one of the benefit of our framework to design new smoothness measures for new problems on hypergraphs.

We conduct experiments to show that our sparsely smooth models are beneficial to learning irrelevant and noisy data, and usually give similar or improved performances compared to non-sparse models. We compare predictive performance on benchmark categorical data with hyperedges being objects having a common categorical value. Experimental results can be found in Figure 2. The highlighted numbers are the highest AUCs, showing the highest performance, which means the most suitable models for the data.

Reference:

Canh Hao Nguyen and Hiroshi Mamitsuka, "Learning on Hypergraphs with Sparsity", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, accepted (2020).

Dataset	n	m	Dense	Hyperedge Select.	Node Select.	Joint Select.
HayesRoth	132	15	0.587±0.044	0.600±0.071	0.758±0.076	0.746±0.067
Lenses	24	9	0.730±0.215	**0.574±0.248**	0.767±0.227	0.770±0.227
Congress	435	48	0.373±0.011	0.473±0.012	0.444±0.010	**0.306±0.034**
Spect	267	44	0.384±0.035	0.400±0.021	0.405±0.057	0.404±0.031
TicTacToe	958	27	0.468±0.009	0.476±0.009	0.481±0.019	0.476±0.009
Car	1728	21	0.692±0.043	**0.462±0.026**	0.748±0.043	0.740±0.044
Monks	124	17	0.469±0.008	0.437±0.023	0.528±0.029	0.504±0.004
Balance	625	20	0.831±0.013	0.955±0.010	0.916±0.014	**0.629±0.044**

Figure 2. Results of newly proposed sparsely smooth models on benchmark data.