Bioinformatics Center
– Bio-knowledge Engineering –

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

Prof
MAMITSUKA, Hiroshi
(D Sc)

Assist Prof
HANCOCK, Timothy Peter
(Ph D)

Assist Prof
KARASUYAMA, Masayuki
(D Eng)

PD
NGUYEN, Hao Canh
(Ph D)

Proj Res
NATSUME, Yayoi
(D Agr)

Students

TAKAHASHI, Keiichiro (D2) MOHAMED, Ahmed (D1) CHEN, Zhuoxin (M1)

Visitors

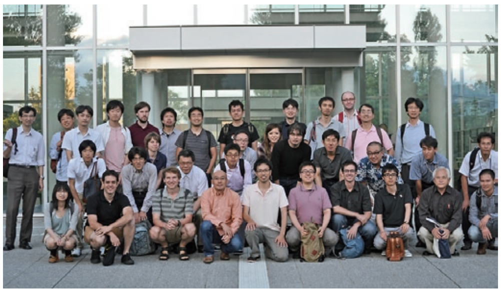
Prof TULLIUS, Tom	Boston University, U.S.A., 11 January
Ms LEE, En-Shiun Annie	University of Waterloo, Canada, 1 February
Assoc Prof ZHU, Shanfeng	Fudan University, China, P.R., 24 March–23 August
Mr ZHENG, Xiaodong	Fudan University, China, P.R., 9 April–7 July
Mr JOHNSTON, Ian	Boston University, U.S.A., 15 May–15 August
Assist Prof CARVALHO, Luis	Boston University, U.S.A., 24 June–7 July
Mr DE LÉSÉLEUC, Sylvain	École Polytechnique, France, 19 July
Dr FUJIMAKI, Ryohei	NEC Labs America, U.S.A., 10 September
Prof WONG, Limsoon	National University of Singapore, Singapore, 1 November

Scope of Research

We are interested in graphs and networks in biology, chemistry and medical sciences, which include 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 developed techniques to real data to demonstrate the performance of the methods and further to find new scientific insights.

KEYWORDS

Bioinformatics
Computational Genomics
Data Mining
Machine Learning
Systems Biology



Selected Publications

Hancock, T.; Mamitsuka, H., Boosted Network Classifiers for Local Feature Selection, *IEEE Transactions on Neural Networks and Learning Systems*, **23(11),** 1767-1778 (2012).

Hancock, T.; Wicker, N.; Takigawa, I.; Mamitsuka, H., Identifying Neighborhoods of Coordinated Gene Expression and Metabolite Profiles, *PLoS One*, **7(2)**, e31345 (2012).

Shiga, M.; Mamitsuka, H., A Variational Bayesian Framework for Clustering with Multiple Graphs, *IEEE Transactions on Knowledge and Data Engineering*, **24(4),** 577-590 (2012).

Nguyen, C. H.; Mamitsuka, H., Latent Feature Kernels for Link Prediction on Sparse Graphs, *IEEE Transactions on Neural Networks and Learning Systems*, **23(11),** 1793-1804 (2012).

Shiga, M.; Mamitsuka, H., Efficient Semi-Supervised Learning on Locally Informative Multiple Graphs, *Pattern Recognition*, **45(3),** 1035-1049 (2012).

Imposing Network Structures for Feature Selection with Omic Data

Networks have become a common place to represent the relationship structure across many variables. For small numbers of variables, networks provide an intuitive model of the structure present within a dataset. However, as the size of the network model increases its representative power diminishes. In an effort to maintain the effectiveness of large network models, feature selection algorithms are employed to extract the relevant structure which is related to a specific phenomena. Currently, the requirement for accurate network feature selection algorithms is essential as the sizes and complexities of the known networks continue to grow.

Supervised classification algorithms are also commonly used as network feature selection methods. The size of the network structures has lead to the development of network regularization algorithms. These methods exploit a sparsity assumption to identify the minimum set of network nodes required to optimize classification performance. When the network structure is large and noisy the sparsity assumption made by these methods is appropriate it will enforce the selection of the minimal set of features required for accurate classification. However some networks in biology, such as metabolic networks, are known to possess highly coordinated responses to external phenomena. These responses potentially activate large sections of the network.

In highly correlated environments a related class of models, ensemble methods such as bagging and boosting, are known to perform well. Ensemble methods seek to represent the structure of a large complex dataset through a combination of small models which are built on a subset of important dataset features. In this research we observed an analogous idea to ensemble methods within factorized network probability distributions. Based upon this similarity we propose two novel optimization algorithms, Boosted Expectation Propagation (BEP) and Boosted Message Passing (BMP) (Hancock and Mamitsuka; 2010, 2012). Neither BEP nor BMP assume a sparse solution, but instead seek a weighted average of all network features where the weights are used to emphasize all features which are useful for classification. In this research we focus on applying BEP and BMP to real world networks. Furthermore, we investigate the similarity in selected features and performance between BEP and BMP and network regularized models. We compare model performances on two different types of biological networks, metabolic networks and protein-protein interaction (PPI) networks using microarray data. Our results on real world networks, presented in Figure 1, show that to extract features from correlated networks the assumption of sparsity will adversely effect classification accuracy and feature selection ability.

References

[1] Hancock, T.; Mamitsuka, H., Boosted Network Classifiers for Local Feature Selection, *IEEE Transactions on Neural Networks and Learning Systems*, **23(11),** 1767-1778(2012).

[2] Hancock, T.; Mamitsuka, H., Boosted Optimization for Network Classification, *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS 2010) (JMLR: Workshop and Conference Proceedings)*, **9,** 305-312, Sardinia, Italy, May 2010, MIT Press.

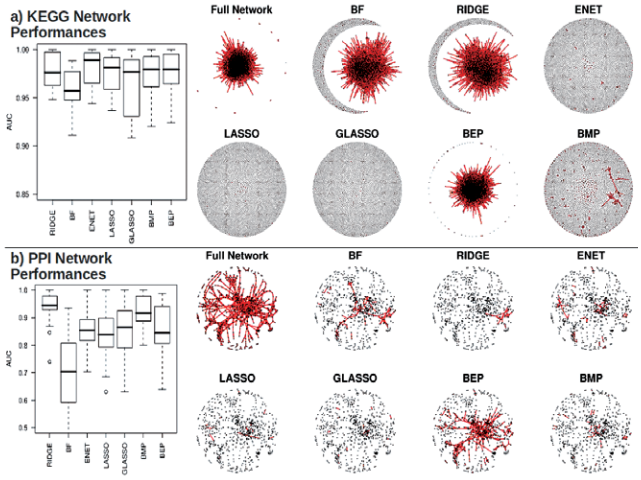


Figure 1. Imposing network structures for feature selection with omic data. The performance (in AUC) and the selected features of the proposed methods, BEP and BMP, and the comparison penalized approaches on the (a) KEGG metabolic network classifying heat stress in yeast, and (b) BIOGRID PPI network classifying tumor occurrence in humans. In the networks the selected features are represented as edges on the network. Good feature selection performance is assumed to highlight connected submodules.