Knowledge Discovery in Biological Big Data using Near Unsupervised Learning

Keynote presentation

Saman K. Halgamuge
Department of Mechanical Engineering and Biomedical Engineering Program
Melbourne School of Engineering, University of Melbourne, Vic 3010, Australia
e-mail: saman@unimelb.edu.au

Keywords-Unsupervised Learning; Big Data; Bioinformatics; Metagenomics; Metabolomics; Neuroengineering

I. SUMMARY

Unsupervised learning is used for analysing and clustering data when the expected cluster labels are completely absent. When we know only a little about the data labels, i.e., class labels are scarce, but available for a small amount of data, it is still challenging to make conclusions, although this may be the case for many real world data mining problems. We name the type of learning algorithms useful in this scenario as Near Unsupervised Learning (NUL). My group has been developing NUL algorithms over a period of 14 years [1,7,11] and some of these developments are based on Growing Self Organising Maps [9-11]. The concept and the algorithm development in NUL and the application in various biological data mining problems will be discussed. Some "unusual" features and signatures captured by my team will also be presented [6]. Real problems attempted using NUL includes the following:

1) Metagenomics involves the challenging problem of clustering and eventually labeling genomic data of microbial species that cannot be easily grown in laboratories [1,5,6,8]. We only know about 2% of these species found on Earth. Could this be the life form we expect to find on another planet? How do we understand and use some unique characteristics of microbes living in our environment and our body?

2) Analysing metabolite profiles of various wheat plants to understand how some type of plants can survive droughts is an area where NUL can provide good solutions [3]

3) Can we analyse signals coming from biological neural networks grown on wet labs to differentiate the sick brain tissues from healthy ones? [2] Our collaborating researchers create mice with brain diseases and analyse the brain tissues with and without the introduction of drugs. Which drugs (for example drugs preventing epileptic attacks) are more effective on a particular type of sickness?

4) Can we find information hidden in data from cancer patients missed by experts, in particular, when the classification of some cancer types are disputed? [4,10]

ACKNOWLEDGMENT

This work is funded by Australian Research Council Grants: DP150103512, LP140100670 and DP1096296, National Health and Medical Research Council Grant 1059665 and YourGene Ltd. The current students: C. Wijetunga, D. Mendis, D. Jayasundara, K. Amarasinghe, previous students K. Chan, Z. Li and D. Alahakoon and collaborators: B. Chang, A. Hsu, I. Saeed, S. L. Tang, J. Li, S. Petrou, A. Stewart, M. Niranjan, U. Roessner, J. Browne, A. Bacic and S. Mahewararajah are acknowledged.

REFERENCES

[1] D. Jayasundara, I. Saeed, S. Maheswararajah, B.C. Chang, S-L. Tang and S. K. Halgamuge, ViQuaS: An improved reconstruction pipeline for viral quasispecies spectra generated by next-generation sequencing, Bioinformatics, Oxford Univ Press, December 2014.

[2] D. Mendis, S. Petrou and S. K. Halgamuge, "Neuromechatronics with In-Vitro Microelectrode Arrays", Chapter in Book edited by C. De Silva, S. K. Halgamuge et. al., 2014, CRC Press.

[3] C.D. Wijetunge, Z. Li, I. Saeed, J. Bowne, A.L. Hsu, U. Roessner, A. Bacic and S.K. Halgamuge, "Exploratory Analysis of High-Throughput Metabolomic Data", Metabolomics, 2013, 9 (6), 1311-1320, Springer.

[4] K. C. Amarasinghe, J. Li and S. K. Halgamuge, "CoNVEX: copy number variation estimation in exome sequencing data using HMM", BMC bioinformatics, Volume 14, Issue Suppl 2, BMC, 2013

[5] CH Tseng, PW Chiang, FK Shiah, YL Chen, JR Liou, TC Hsu, S. Maheswararajah, I. Saeed, S. K. Halgamuge and S. L. Tang, "Microbial and viral metagenomes of a subtropical freshwater reservoir subject to climatic disturbances", The ISME journal, Nature Publishing Group, 2013

[6] I. Saeed, S.L. Tang and S. K. Halgamuge, "Unsupervised discovery of microbial population structure within metagenomes using nucleotide base composition", Nucleic Acids Research, Volume 40, Issue 5, 2012, Oxford University Press.

[7] A. L. Hsu and S. K. Halgamuge, "Class structure visualization with semi-supervised growing self-organizing maps", Neurocomputing, Vol: 71 Issue: 16-18, Pages: 3124-3130, Elsevier, 2008.

[8] C.K. Chan, A.L. Hsu, S.K. Halgamuge and S.L. Tang, "Binning Sequences Using Very Sparse Labels within A Metagenome", BMC Bioinformatics, 2008, 9:215, 28 April 2008.

[9] S. M. Guru, A. Hsu, S.K. Halgamuge and S. Fernando, An Extended Growing Self-Organising Map For Selection of Clustering in Sensor Networks, International Journal of Distributed Sensor Networks, Taylor & Francis, Vol 1, No 2, 2005

[10] A. Hsu and S. Tang and S. K. Halgamuge, An Unsupervised Hierarchical Dynamic Self-Organising Approach to Cancer Class Discovery and Marker Gene Identification in Microarray Data, Bioinformatics, Oxford University Press, November 2003.

[11] D. Alahakoon, S. K. Halgamuge, and B. Srinivasan. Dynamic Self Organising Maps with Controlled Growth for Knowledge Discovery, IEEE Transactions on Neural Networks, May 2000