# IMECS 2015 Keynote Speech III:

# Big Data Analytics using Unsupervised Learning: Can Subsample Error Graphs reveal more than Clustering?

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- Clustering, Big Data Analytics; Bioinformatic, Metagenomics, Metabolomics***

## I. SUMMARY

In collaboration with researchers in Academia Sinica and Metabolomics Australia/Department of Botany at Melbourne, we have been working in two areas of Bioinformatics: Metabolomics focusing on microbes and Metagenomics focusing on plants [1-4]. Profiling large sets of data resulted from technological advances in whole genome sequencing and MALDI Imaging type technologies that can reveal vital information about the environment and plants, which is our major or primary source of food on Earth. Recently we have demonstrated considerable success in using unsupervised clustering techniques to analyse genetic and metabolomic data. This includes analysis of viral quasi species [1], drought resistance in wheat [2] and microbial metagenomes [3] [4].

Some microbes in the environment appear to look very similar and found "living together" in communities in non - separable ways, making them harder to culture in a lab. To make matters worst, considering our belief, if it is correct at all, that we know only about up to 2% of the Microbes around us. When we know only so little about the data labels, in this case, about the identity of the species, it is challenging to recognise patterns associated with the genomes of the separate species. In 2013, Isaam Saeed of the presenter's research group suggested a new method [3] to separate Microbial genomes based on the Sample error gradient and he called this method "Oligonucleotide Frequency Derived Error Gradient (OFDEG)" which seems to characterise microbial genomes very well. The calculation of OFDEG is based on the errors produced when data is subsampled at varying subsample sizes. Since the underlying mathematical concept of OFDEG lies in subsampling, which is not restricted to metagenomics, we are extending its usage into other domains. We also observed that are parameters other than gradient that can be used. Therefore, we call this class of methods, Sab-sample Error Graphs (SEGs). The mathematics behind it's success, particularly, in Metagenomics and potentially also in Metabolomics, made us further explore and compare this new method against previously successful Unsupervised clustering methods.

We will present the concept of SEC and discuss its use in comparison to near Unsupervised learning methods. We will draw examples from on-going Metabolomics and Metagenomics research of the author's group and others.

## ACKNOWLEDGMENT

This work is funded by Australian Research Council Grants: "Near Unsupervised computational methods for exploring omic data (DP150103512), "Discovering Patterns using Near Unsupervised Learning to Support the Quick Detection of New Animal Disease Outbreaks Caused by Viruses" (LP140100670) and YourGene Australia. The current students: C. Wijetunga, D. Mendis, D. Jayasundara, K. Amarasinghe, previous students Z. Li, U. Premaratne, K. Chan and collaborators: S. L. Tang, B. Chang, A. Hsu, I. Saeed, U. Roessner, J. Browne and A. Bacic 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] 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.

[3] 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.

[4] 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.

## BIOGRAPHY OF PROFESSOR SAMAN HALGAMUGE

Prof. Saman Halgamuge is a Professor of the Department of Mechanical Engineering and the school wide initiative on Biomedical Engineering and Associate Dean (International) for the Melbourne School of Engineering, The University of Melbourne. He graduated with Dipl.-Ing and PhD degrees in Electrical Engineering at Technical University of Darmstadt, Germany.

His research interests are in Big Data Analytics and Optimization and in particular bio-inspired methods focusing on applications in Mechanical Engineering and Bioengineering. These applications vary from Sensor Networks in Irrigation, Smart Grids, and Sustainable Energy generation to Bioinformatics and Neuro-Engineering. Since arriving in Australia in 1996, Professor Halgamuge has obtained research grants totaling over $8.3 million. These grants include funds worth $4.3 million from 17 Australian Research Council grants (4 Discovery projects and 8 Linkages) and a NHMRC project grant, local and European industry, contracts and grants from other research funding agencies and large-scale ARC network and infrastructure grants worth about $4 million.

He has completed supervision of 27 PhD students and currently supervises a group of 13 PhD students. He is an Associate Editor of BMC Bioinformatics and founding co-editor of Frontier’s journal on Mechanical Engineering- Mechatronics section. He published over 250 research papers including a research book, 5 edited books, 20 book chapters, 90 journal articles, and over 130 refereed conference papers attracting 5100 Google Scholar (h-factor: 29) and 2300 ISI citations. He is listed among the top 1% most cited researchers by ISI's ESI. His full profile is at http://scholar.google.com.au/citations?sortby=pubdate&hl=en&user=9cafqywAAAAJ&view_op=list_works