IITB-Monash Research Academy

Indian Institute of Technology Bombay | MONASH University

An Indian-Australian research partnership

Project Title:	A new paradigm that enables big data mining	
Project Number	**IMURA0394**	
Monash Supervisor(s)	Prof. Kai Ming Ting Prof. Geoff Webb Prof. David Albrecht	*Full names and titles*
Monash Primary Contact:	KaiMing.Ting@monash.edu	*Email, phone*
Monash Head of Department:	Prof. Graham Farr	*Full name, email*
Monash Department:	Clayton School of IT	*Full name*
Monash ADRT:	Kai Ming Ting	*Full name, email*
IITB Supervisor(s)	Prof. Ganesh Ramakrishnan	*Full names and titles*
IITB Primary Contact:	ganesh@cse.iitb.ac.in	*Email, phone*
IITB Head of Department:	S Sudarshan	*Name, Email,*
IITB Department:	Computer Science and Engineering	*Full name*

Research Academy Themes:

Highlight which of the Academy's Theme(s) this project will address?

(*Feel free to nominate more than one. For more information, see www.iitbmonash.org)*

1. Advanced computational engineering, simulation and manufacture
2. Infrastructure Engineering
3. Clean Energy
4. Water
5. Nanotechnology
6. Biotechnology and Stem Cell Research

The research problem

Define the problem

In this project, we establish a new paradigm for data mining – a mathematical framework that has a core data modelling mechanism called **mass estimation**which is fundamentally different from existing paradigm:density estimation. Mass estimation delineates the *centrality* of a data cloud whereas density estimation describes the *compactness* of a data cloud.

Mass estimation has sublinear runtime and memory space requirements and can potentially be reduced to constant, irrespective of the input data size. This means that it can be computed significantly faster than density estimation, using a significantly smaller memory space. In addition, it can approximate

similarity between two instances without the use of a similarity metric. These features make mass estimation a better modelling mechanism than density estimation to solve big data problems, especially in the context of data streams and high dimensional problems.

The project builds upon the research achieved in this breakthrough to yield a principled theory and to address the hard problems in big data mining, specifically data streams and high dimensional problems. The project has two goals:

Goal 1. Build a theory of mass estimation.

Goal 2. Create mass-based algorithms that make mining big data with high dimensionality a reality

In the nascent data-centred economy, "data are becoming the new raw material of business: an economic input almost on a par with capital and labour" [Data, data everywhere. *The Economist*. 25 February 2010.]. The new paradigm enables the expanding vast amount of data to be mined effectively, efficiently and timely that would otherwise impossible in the existing paradigm. This unlocks not only the hidden information in big data, but also economic values worth billions of dollars.

This project

- Builds new methods based on mass, and pioneers the application in data mining areas such asclassification, clustering, anomaly detection, information retrieval in big data with high dimension and in data streams.
- Creates mass-based similarity measures that have distinctive properties in comparison with existing distance-based similarity measures and non-metrics.

A number of algorithms based on mass estimation, created up to date, are available in public domain:

1. iForest: an anomaly detection algorithm. http://sourceforge.net/projects/iforest/
2. Mass estimation and its associated algorithms for clustering and classification. http://sourceforge.net/projects/mass-estimation/

Project aims

Define the aims of the project

Expected outcomes

Highlight the expected outcomes of the project

How will the project address the Goals of the above Themes?

Describe how the project will address the goals of one or more of the 6 Themes listed above.

Capabilities and Degrees Required

List the ideal set of capabilities that a student should have for this project. Feel free to be as specific or as general as you like. These capabilities will be input into the online application form and students who opt for this project will be required to show that they can demonstrate these capabilities.

Proficiency in programming in either C, Java, or Matlab.

Have background in algorithmic methodology in data mining or machine learning (but not a must)