Editorial

Naveen Dahiya*

Department of Computer Science and Engineering,
Maharaja Surajmal Institute of Technology,
C-4, Janakpuri, New Delhi, 110058, India
Email: naveendahiya16@yahoo.com
*Corresponding author

Vishal Bhatnagar

Department of Computer Science and Engineering,
Ambedkar Institute of Advanced
Communication Technologies and Research,
Geeta Colony, Delhi, 110031, India
Email: vishalbhatnagar@yahoo.com

Nilanjan Dey

Department of Information Technology,
Techno India College of Technology,
Kolkata, West Bengal, 700156, India
Email: neelanjan.dey@gmail.com

Biographical notes: Naveen Dahiya received his BE in Computer Science and Engineering from the Maharshi Dayanand University, Rohtak, Haryana, India, in 2003, MTech in Computer Engineering from the Maharshi Dayanand University, Rohtak, Haryana, India, in 2005 and PhD from the YMCA University of Science and Technology, Faridabad, Haryana, India, in 2016. He is working as an Associate Professor and Head in Computer Science and Engineering Department at the Maharaja Surajmal Institute of Technology, C-4, Janak Puri, New Delhi, India. His research interests include database systems, data warehouse and data mining.

Vishal Bhatnagar holds BTech, MTech and PhD in the engineering field. He has more than 20 years of teaching experience in various technical institutions. He is currently working as a Professor in Computer Science from the Engineering Department at the Ambedkar Institute of Advanced Communication Technologies and Research (Government of Delhi), GGSIPU, Delhi, India. His research interests include database, advance database, data warehouse, data-mining, social network analysis and big data analytics. He has to his credit more than 100 research papers in various international/national journals and conferences.

Nilanjan Dey is an Assistant Professor from the Department of Information Technology at the Techno India College of Technology (under Techno India Group), Kolkata, India. He is also a Visiting Professor at the Duy Tan University, Vietnam, He was an Honorary Visiting Scientist from the Global Biomedical Technologies Inc., CA, USA (2012–2015). He is a Research

Scientist of the Laboratory of Applied Mathematical Modeling in Human Physiology, Territorial Organization of Scientific and Engineering Unions. He is an associated member of the University of Reading, London, UK and scientific member of Politécnica of Porto. He was awarded with PhD from the Jadavpur Univeristy, in 2015. His main research interests include medical imaging, machine learning, computer aided diagnosis as well as data mining.

The information and communication technology (ICT) related exponential growth has increased the demand for big data analytics (BDA). BDA involves the handling of a gigantic data for storage and investigation. The evolving field of BDA owns many challenges in various fields including drug delivery, healthcare, surveillance, weather forecasting, etc. In recent years, there is a proliferation in the amount of data generated.

Big data is a current and hot topic of IT industry. Harnessing such huge amount of data is a tough job and thus it requires business intelligence (BI) or analytics. BI is required to explore new knowledge, relationships and patterns among different data. Big data has already become a prominent part of a $64 billion database and analytics souk. Gartner has defined big data by giving three characteristics (volume, velocity, variety) prevalently known as three V's of big data. Later on, IBM defined one more characteristics (veracity) and provides the theory of four V's. It also includes the three V's given by Gartner. Nowadays organisations are struggling with fifth V of big data, finding value contained in big data. Analytic souk is leaving no stone unturned to tame the available data that yields interesting results and better insights of industry or organisation.

- Volume: It refers to the amount of data. When data size surpasses from terabytes to exabyte's, it refers to big data. Big firms like facebook, twitter generate billions of data daily and uses BDA to retrieve valuable information when require.
- Velocity: The speed at which data is received and processed. It corresponds to data in motion. Real-time applications and internet of things (IOT) requires abrupt response. So in such cases, time for harnessing the data should be in seconds to milliseconds.
- Variety: This feature depicts different formats of data. Big data can handle any form of data whether structured, semi-structured or unstructured.
- Veracity: This refers to data in doubt. Veracity in data analytics is a non-desired feature in data. It is the uncertainties and noise present in the data.
- Value: Characteristic refers to the intrinsic value contained in big data.

This special issue is a collection of the seven papers which are written by eminent professors and researchers from different countries. The papers were initially peer reviewed by the editorial board members and reviewers who themselves span over many countries.

- Paper 1: 'An ensemble clustering method for intrusion detection', by Kapil K. Wankhade and Kalpana C. Jondhale

  This paper mainly focuses on detection rate and false alarm rate so to resolves these problems a hybrid method, ensemble clustering has been proposed. This method tries

to increase detection rate with lowering false alarm rate. The method has been tested on KDDCup'99 network intrusion dataset and performs well as compared with other algorithms in terms of detection rate false alarm rate.

- Paper 2: 'Empirical investigation of dimension hierarchy sharing-based metrics for multidimensional schema understandability', by Anjana Gosain and Jaspreeti Singh

  Over the last years quality has gained lot of importance in the development of data warehouse systems. Predicting understandability of multidimensional schemas could play a key role in controlling data warehouse quality at early stages of development. In this area, some effort has been spent to define structural metrics and identity models for assessing quality of these systems. Of the structural properties used to define metrics, aspects of dimension hierarchies and its sharing plays primary role to enhance analytical capabilities of multidimensional schemas, thereby affecting their quality. The authors have previously proposed structural metrics based on aforementioned aspects. The objective of this study is to apply principal component analysis (PCA) to find whether our metrics are improvements over the other existing metrics; and to apply logistic regression to study whether the metrics (selected as relevant in the extracted principal components) combined together are indicators of multidimensional schema understandability. The results of PCA confirm that our structural metrics based on the concept of sharing are different from other such metrics existing in the literature. Further, the metrics selected as principal components can be used in combination to predict understandability of data warehousemultidimensionalschemas.

- Paper 3: 'Detecting concept drift using HEDDM in data stream', by Snehlata S. Dongre, Latesh G. Malik and Achamma Thomas

  In evolving data stream, when its concept undergoes a change it is known as concept drift. Detecting concept drift and handling it is a challenging task in data stream mining. If an algorithm is not adapted to concept drift, then it directly affects its performance. A number of algorithms have been developed to handle concept drift, but they are not suited for both – sudden concept drift and gradual concept drift. Thus, there is a demand for an algorithm that can react to both the types of concept drift as well as incur less computational cost. A new approach – hybrid early drift detection method (HEDDM) – has been proposed for drift detection, which works with an ensemble method to improve the performance.

- Paper 4: 'Dynamic social network analysis and performance evaluation', by Sanur Sharma and Anurag Jain

  Social media in today's age is on a tremendous increase in terms of its usage and the enormous amount of data it generates which includes personal details of users, their images and the content that is being shared on such open source platforms. This has led to a lot of research and analysis of such networks and data that exists in social media. This paper is focused on dynamic analysis of social networks, where snapshots of network are taken at regular intervals and are analysed on various performance measures. The real time e-mail dataset of a company (ENRON) has been evaluated and visualised dynamically. The network measures are evaluated at each timestamp and clustering is performed on that data and its performance is

calculated on various measures. Tabu search optimisation algorithm has been used for clustering the time stamped data and a comparison is done between the fixed size cluster and variable size clusters. The results suggests that for certain time stamps the value of precision, recall and f measure for fixed size clusters are better than the variable size clusters. These measures can further be used for the selection of the dynamic clustering techniques for social network analysis.

- Paper 5: 'Measuring harmfulness of class imbalance by data complexity measures in oversampling methods', by Anjana Gosain, Anju Saha and Deepika Singh

  Many real world applications consist of skewed datasets which result in class imbalance problem. During classification, class imbalance cause underestimation of minority classes. Researchers have proposed a number of algorithms to deal with this problem. But recent research studies have shown that some skewed datasets are unharmful and applying class imbalance algorithms on these datasets lead to degenerated performance and increased execution time. In this research paper, we have pre-estimated the degree of harmfulness of class imbalance for skewed classification problems, using two of the data complexity measures: scatter matrix-based class separability measure and ratio of intra-class versus inter-class nearest neighbours. Also the performance of oversampling-based class imbalance classification algorithms have been analysed with respect to these data complexity measures. The experiments are conducted using k-nearest neighbour (*k*-nn) and naive Bayes as the base classifiers for this study. The obtained results illustrate the usefulness of these measures by providing the prior information about the nature of the imbalance datasets that help us to select the more efficient classification algorithm.

- Paper 6: 'Threshold-based empirical validation of object-oriented metrics on different severity levels', by Aarti, Geeta Sikka and Renu Dhir

  Software metrics has become desideratum for the fault-proneness, reusability and effort prediction. To enhance and intensify the sufficiency of object-oriented (OO) metrics, it is crucial to perceive the relationship between OO metrics and fault-proneness at distinct severity levels. This paper characterise on the investigation of the software parts with higher probability of occurrence of faults. We examined the effect of thresholds on the OO metrics and build the predictive model based on those threshold values. This paper also instanced on the empirical validation of threshold values calculated for the OO metrics for predicting faults at different severity levels and builds the statistical model using logistic regression. This paper depicts the detection of fault-proneness by extracting the relevant OO metrics and focus on those projects that falls outside the specified risk level for allocating the more resources to them. We presented the effects of threshold values at different risk levels and also validated results on the KC1 dataset using machine learning and different classifiers. The results evaluated on the receiver and operator (ROC) parameters concluded that threshold methodology has great potential for conducting prediction of faults and shows that analysis of result using machine learning techniques outperforms as compared to logistic regression.

We hope that this special issue will provide a quality publication with innovative ideas and implementation methodology to upcoming buddy researchers and users in the modern day era. This will also pave ways for future innovative ideas in this field.

The guest editors would like to thank all the authors for submitting their manuscripts in this special issue and acknowledge the reviewers for their valuable contributions in reviewing the papers and providing constructive and useful comments to the authors. Finally, the guest editors would like to specially thank the Editor-in-Chief of *Int. J. Intelligent Engineering Informatics* (*IJIEI*), Professor Ahmad Taher Azar (Benha University, Egypt) for his great help and support in organising and coordinating the publication of this special issue.