Introduction

Jia Zeng*

Baylor College of Medicine,
Texas Medical Center,
Houston, Texas, USA
Email: jzeng@bcm.edu
Email: increasezj@gmail.com
*Corresponding author

Mehmet Tan

Department of Computer Engineering,
TOBB University of Economics and Technology,
Ankara 06560, Turkey
Email: mtan@etu.edu.tr
Email: mehmet.tan@gmail.com

Biographical notes: Jia Zeng received her PhD degree from University of Calgary in 2009. She is currently a Research Associate at the Department of Biochemistry & Molecular Biology at Baylor College of Medicine. Her research interests include computational biology, machine learning and cancer epigenetics.

Mehmet Tan received BSc, MSc and PhD degrees in Computer Engineering from Middle East Technical University, Ankara, Turkey in 2000, 2003 and 2009 respectively. He was a visiting scholar in University of Calgary from Feb. 2007 to Nov. 2008. He is currently an Assistant Professor in Department of Computer Engineering, TOBB University of Economics and Technology, Ankara, Turkey. His research interests include machine learning and data mining with a special focus on bioinformatics and cheminformatics.

The past two decades have seen an explosion of information that is made readily available for annotation and analysis. The increase in the amount of data presents a challenge for their efficient utilisation. The series of the *Information Reuse and Integration* (IRI) conferences focuses on the topics of optimising information utilisation by encouraging data reuse and integration. Such an effort is applicable and useful to a wide range of applications in many different industries. This special issue presents six papers. One has been selected from the papers which were submitted in response to a call for papers which was distributed at large. And five papers have been chosen from the papers that were accepted and presented at the 2011 *IEEE International Conference on Information Reuse and Integration* (*IEEE IRI'2011*). Based on the outstanding reviews given by the referees of IRI'11, we invited the authors of these five papers to prepare an extended version of their conference paper to be considered for inclusion in this special issue. All of this extension was considered significant and original based upon the feedback from a second round of reviews conducted by the special issue review board. Therefore, we proudly feature these six papers in our special issue.

Papers in this special issue cover several different aspects of the main theme of the IRI conferences. The first two papers investigate the analysis and processing of the information obtained from the internet. Another two papers delve into classification and clustering respectively which are the classic topics of data mining. The third group of papers revolves around the theme of feature selection and feature ranking.

Specifically the first paper, authored by Nagi et al., proposes a robust framework for recommending the restructuring of websites that helps the site owner to increase the number of satisfied visitors by allowing the visitors to have the flexibility to reach their target pages. The proposed framework heavily utilises the web log data to investigate the usage patterns leading to recommendation of the most appropriate way to link pages of a website. Thanks to the application of an array of advanced machine learning algorithms such as genetic algorithm based clustering, frequent pattern mining and network analysis techniques, the said framework identifies the pages that are frequently accessed and/or clustered together and then links these pages in order to allow for faster access to the information deemed relevant by a certain group of users.

The second paper by Kianmehr and Koochakzadeh explores the topic of cybercrime prediction and present an alternative solution to the current intrusion detection system by examining the socio-economic characteristics of IP geo-locations. This is done via the use of the IP address of a web service request which then identifies the physical location from where the request was sent. The socio-economic attributes of people living in that particular area are then collected and used to represent the characteristics relevant to the potentiality and seriousness of a cybercrime associated with a service request. Then a classification algorithm can be employed to establish a prediction model. The authors conducted a case study which investigated different types of cybercrime ratios and a set of attributes about population distribution based on age, race, gender and housing status. They concluded that these socio-economic characteristics of the geo-location of a service request are indeed important for making appropriate prediction of the cybercrimes.

The third paper by Zhu, Lin and Shyu focuses on examining the necessary pre-processing steps for many classification models. They presented a novel discretisation algorithm based on correlation maximisation using multiple correspondence analysis which is an effective technique to capture the correlation between multiple variables. The discretised feature not only produces a concise summarisation of the original numeric feature but also provides the maximum correlation information to predict class labels. Results from extensive experiments demonstrate that the proposed algorithm can automatically generate a set of features that produce the best classification results on average.

The fourth paper authored by Salunke, Liu and Rege investigates the topic of co-clustering which derives sub-matrices of the data matrix by simultaneously clustering the data instances and features of the matrix. The authors proposed a novel semi-supervised constrained co-clustering algorithm with non-negative matrix factorisation to integrate domain knowledge in the forms of must-link and cannot-link constraints. Two approaches for integrating domain knowledge were explored including a distance metric learning approach and information theoretic learning approach. The proposed paradigm was evaluated in web-service community discovery and text mining.

The fifth paper by Altidor, Khoshgotaar and Napolitano investigates the stability performance of 17 filter-based feature ranking techniques. The authors focused on the scenarios where noise is widely present and demonstrated that some feature ranking

techniques are inherently more sensitive to noise than others. Given the prevalence of noise in real-world data, the research findings presented in the paper offer some useful insight which can guide the choice of feature ranking scheme in the context of noisy data.

The last featured paper, authored by Shanab, Khoshgoftaar, Wald and Hulse, looks into the evaluation of the importance of data preprocessing order when combining feature selection and data sampling. Two challenging problems in the field of machine learning are class imbalance and high dimensionality. The authors propose three different approaches for addressing both problems simultaneously. All these three approaches combine sampling and feature selection. A thorough analysis was performed and a conclusion was drawn regarding the performance of each approach.

Finally, we would like to thank all the authors for their efforts in refining and extending their papers for this special issue. The Editor-in-Chief, David Taniar's guidance and support are particularly appreciated. Thanks also go to authors, reviewers and production department at Inderscience for their assistance in making this special issue published in a timely fashion.