

Special issue on soft computing for big data and social informatics

Chun-Hao Chen¹ · Chuan-Kang Ting² · Yusuke Nojima³

Published online: 11 May 2017
© Springer-Verlag Berlin Heidelberg 2017

Welcome to the special issue on soft computing for big data and social informatics. With rapid advance in wireless techniques and mobile devices, big data and social informatics have received considerable attention in recent years. The big data appears in various areas like engineering, commerce, finance, medical science and social informatics involves the study of social aspects of computerization, including the role of information and communication technology in social and organizational change, and the use of it in social contexts. The objective of this special issue is to highlight an ongoing research, bring forward recent advances and present state-of-the-art developments on soft computing for big data and social informatics in various domains. There are totally ten articles included, mainly selected from The Fifth ASE International Conference on Big Data (ASE BIGDATA2015) and The Fourth ASE International Conference on Social Informatics (SOCIALINFORMATICS2015), which were jointly held in Kaohsiung, Taiwan on October 7 to 9, 2015. A few of them were selected from the 2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI2015), which was held in Tainan, Taiwan. The contents of the articles are briefly introduced below.

The first two papers are on clothing image recognition and image annotation for large-scale data. The first work authored by Chen et al. deals with visual-based clothing image recognition using deep net architectures. The authors explore convolutional neural networks with a goal of resolving clothing style classification and retrieval tasks. To reduce training

complexity, low-level and mid-level features were learned in the deep model on large-scaled datasets, and transfer learning is then incorporated by the fine-tuning pre-trained model using the clothing dataset. Since a large amount of collected data need huge computation for tuning parameters, one architecture inspired from AdaBoost is designed for multiple deep nets trained with a sub-dataset. In addition, to increase system flexibility, two architectures with multiple deep nets with two outputs are proposed for binary-class classification.

In the second paper, Chien et al. consider relationships of crossing images and texts including image-to-text, text-to-text and image-to-image association for large-scale image annotation since managing large-scale image data becomes an important research issue due to considerable increase of digital images in recent years. They first propose the cross image–text annotation framework, on which a set of hybrid learning models are developed and implemented by means of image classifiers, similarity image matching and association mining of image labels. Experiments investigate the performance of the cross image–text framework by evaluating the effectiveness of different annotation models, including individual models, bi-hybrid models and the all-hybrid model, and indicate the hybrid models combining the relationships with both images and texts boost the effectiveness of annotation.

The next two papers are about utility itemset mining and fuzzy–data mining from large transaction data. In the third paper, Lin et al. present an efficient algorithm, named mining uncertain high-utility itemsets (MUHUI), to efficiently discover potential high-utility itemsets (PHUIs) in uncertain data. Based on the probability-utility-list (PU-list) structure, the MUHUI algorithm directly mines PHUIs without generating candidates and can avoid constructing PU-lists for numerous unpromising itemsets by applying several pruning strategies, which greatly improve its performance. In

✉ Chun-Hao Chen
chchen@mail.tku.edu.tw

¹ Tamkang University, New Taipei City, Taiwan

² National Chung Cheng University, Chia-yi County, Taiwan

³ Osaka Prefecture University, Osaka, Japan

the fourth paper, since the increase in data poses serious challenges to GA in the effectiveness and efficiency of finding appropriate membership functions, especially in big data analytics, Ting et al. proposes a GA for enhancing genetic-fuzzy mining of association rules. First, they design a novel chromosome representation considering the structures of membership functions. The representation facilitates arrangement of membership functions. Second, the study presents two heuristics in the light of overlap and coverage for removing inappropriate arrangement. The experimental outcomes also validate the high capability of the proposed GA in genetic-fuzzy mining of association rules.

In the big data era, various data types are formed from research fields. In the following four papers, algorithms are presented and applied on those data types to deal with different specific problems. In the fifth paper, Wu et al. use trading mechanisms to investigate large future data and its implication to market trends. They apply the stop-loss and stop-profit mechanisms to verify the market trends based on two new strategies. Then, they back-test these two strategies on the Taiwan Stock Exchange Capitalization Weighted Stock Index Futures (TAIEX Futures). They compare the numerical results of profits and losses through various stop-loss thresholds and stop-profit thresholds and verify the existence of the momentum effect via applying these two new trading strategies. In addition, they also analyze the market trends through the repeated simulations of random trades with the stop-loss and stop-profit mechanisms. At last, they propose a technique to quantify the momentum effect of a financial market by using Jensen–Shannon divergence.

To recognize cyber threats from open threat intelligence, Lee et al. in the sixth paper design the Sec-Buzzer which is a web-based service not only finding various emerging topics of cyber threats and its corresponding annotations but also providing the possible remedy solutions. The Sec-Buzzer leverages different kinds of open sources, Twitter data and domain-specific blogs and benefits a lot from the community-oriented filtering strategy as well as novel topic-association graph. Therefore, a set of highly contributing Twitter users are grouped and scored as an expert community and information from that will be explored and then efficiently exploited. Demonstrations show that Sec-Buzzer indeed uncovers unseen valuable domain experts to be information providers and identifies emerging topics.

In a high-dimensional microarray gene expression cancer database, using clustering analysis for identifying cancer types is a difficult task because of high-dimensional genes with noise. In the seventh paper, Kannan et al. thus present

effective fuzzy c -means by incorporating the membership function of fuzzy c -means, the typicality of possibilistic c -means approaches and normed kernel induced distance to find cancer sub-types in a microarray gene expression cancer database. The proposed approach successfully finds cancer sub-types in a microarray gene expression cancer database.

In the eighth work, Tsai et al. propose a wharf-based genetic algorithm for scheduling public berths with the aim of reducing reliance on communications and shortening the waiting time of vessels. The proposed algorithm uses a special wharf-based sequential type of chromosomes that keep all of the generated schedules as feasible solutions. In addition, combining the MapReduce technique, the proposed algorithm is able to handle a great number of vessels. Experimental results demonstrate the effectiveness of the proposed algorithm at assigning vessels to appropriate berths as soon as they arrive.

The last two papers of the special issue are about modeling. Since patents are business and financial assets which can enhance a company's competitive position, patent analysis is important for defining business strategies and supporting decision making in organizations. In the ninth paper, Chang et al. design a patent quality classification model using an artificial immune system. They apply the model to predicting the quality of radio-frequency identification (RFID) patents. Using a simple definition of quality, they define each patent's data as an antigen and then compute the affinities of the target patent to all the immune network. If the affinity is larger than a given threshold, the antibody is cloned to the related immune network. After the immune networks are constructed, they exhibit high affinity to the target patent.

In the last paper, Chou et al. propose a context-aware propagation method on Chinese ConceptNet for sentiment polarity prediction. They represent contexts using the LDA topic model by generating a topic for each context. They can then assign concepts different sentiment values for different topics when sentiments are propagated on Chinese ConceptNet. Experiments on both microblog posts and drama dialogue subtitles show that the context-aware approach improves the accuracy of sentiment polarity prediction.

We hope the special issue can bring some interesting ideas and recent advances in the soft computing for big data and social informatics. At last, we are grateful to all the authors for their contributions and to the referees for their vision and efforts. We would also like to express our thanks to Professors Antonio Di Nola (Editor-in-Chief) and Vincenzo Loia (Co-Editor-in-Chief) of Springer Soft Computing Journal, for their great support to realize the special issue.