Intelligent Data Analysis 22 (2018) 1185–1187
DOI 10.3233/IDA-180894
IOS Press

Editorial

Dear Colleague:
Welcome to volume 22(6) of Intelligent Data Analysis (IDA) Journal.

This issue of the IDA journal, the sixth and the last issue of our twenty second year of publication, contains 12 articles representing a wide range of topics related to the theoretical and applied research in the field of Intelligent Data Analysis.

The first three articles are about feature selection and data sampling in IDA. Ma *et al.* in the first article of this issue propose a feature selection method using forest optimization algorithm which is based on contribution degree. Their proposed method uses a contribution degree strategy which is embedded in forest optimization algorithm. The proposed algorithm is verified on some data sets from the UCI repository and their experiments show that the proposed method improves the classification accuracy compared with some other methods. Belouch *et al.* in the second article of this group explain that cloud computing has become a significant trend for the delivery of IT business services, and represents a potential technology resource choice that offers cost effective and scalable processing. The authors argue that in previous research, feature selection has revealed its importance in the recognition of irrelevant and redundant features, which increases detection rates and decreases processing speeds toward the evaluation of intrusive patterns, while reducing computational complexity. The authors propose a Hybrid Filter-Wrapper Feature Selection HFWFS method for DDoS detection, which takes advantage of both filter and wrapper methods, to identify the most irrelevant and redundant features in order to form a reduced input subset. The results of their experiments indicated that the proposed model may reduce the number of features from more than 40 to nine, while maintaining high detection accuracy, in contrast to well-known feature selection methods. Zhou *et al.* in the third article of this issue argue that LDA class of algorithms attempt to improve the time and space efficiency of its Gibbs sampling. However, there could be some limitations since the amount of recycled computation is limited. The authors propose a new algorithm that first rearranges the tokens within one text according to the word types so that the tokens of the same word type are aggregated together and then recycle more computation while making no approximation and ensuring the exactness. The article includes detailed theoretical explanations and comparative experimental analyses on the correctness, exactness and time-efficiency of the proposed approach.

The next two articles are about various forms of data analysis in frequent item sets. Liu *et al.* in the first article of the second group focus on preserving the sensitive utility and frequent itemsets, and present a sanitization approach in which sensitive itemsets are hidden by reducing their support or utility below the minimum thresholds. The authors compare their approach with the state of the art algorithms on various databases. Their experimental results show that the proposed idea outperforms the other algorithms in minimizing the side effects on non-sensitive knowledge and maintaining the database quality after the sanitization process. Zheng *et al.* in the next article on the same topic of frequent item sets propose a new graph mining algorithm that is capable of locating all frequent induced subgraphs in a given set of directed networks. The authors present an incremental coding scheme for representing the canonical form of a graph, evaluate its properties, and develop new techniques for pattern generation suitable for directed networks. Their experimental results based on synthetic data and gene regulatory networks show the performance of their algorithm, and its application in network inference.

The third group of articles in this issue are about various forms of learning in IDA. Denisiuk and Grabowski in the sixth article of this issue present a new c-means clustering algorithm for combined continuous-nominal data which is based on using spherical representation of nominal data. In their approach, the impact of specific features is modelled with corresponding weights in metric definition. A series of numerical experiments on real and synthetic data show that their proposed algorithm can successfully cluster raw, non-normalized data. Zhour and Qi in the seventh article of this issue argue that modelling knowledge as OWL ontologies is an error-prone process, where logical errors or contradictions would be imported. They further argue that, it is almost impossible to manually find errors occurring in large-scale ontologies. In order to tackle this problem, the authors propose a method to find justifications independent from the utilized reasoner. The authors give several optimizing strategies and prove the correctness of their method. Their experimental results show that their method is practical and performs better than existing methods. Gao *et al.* in the last article of this group argue knowledge bases that are generated from Semantic Web lack expressive schema information, especially disjointness axioms and subclass axioms. This makes it difficult to perform many critical Semantic Web tasks like ontology reasoning, inconsistency handling and ontology mapping. The authors propose a novel framework to automatically obtain disjointness axioms and subclass axioms from incomplete semantic data. Their experimental evaluation shows promising results over several real-life incomplete knowledge bases like DBpedia and LUBM by comparing with existing relevant approaches.

And finally, the last group of articles in this issue are about enabling techniques and novel methods. Nogueira *et al.* in the ninth article of this issue address the problem of correctly predicting the illness path in various patients of Acute Kidney Injury by studying different methodologies to predict this disease and propose new distinct approaches based on the idea of improving the performance of the classification. Through the comparison of five different approaches (Markov Chain Model ICU Specialists, Markov Chain Model Features, Markov Chain Model Conditional Features, Markov Chain Model and Random Forest), the authors have concluded that the application of conditional probabilities to this problem produces a more accurate prediction, based on common inputs. Liu *et al.* in the next article of this issue address the sparsity problem in recommender systems in real world social network applications, where side information could be rendered. The authors propose a novel hierarchical Bayesian model named collaborative social deep learning (CSDL), which jointly handles deep learning for the content information and collaborative filtering for general users' following actions, the social network of celebrities and that of general users. Their experiments on two real-world datasets show the effectiveness of their proposed model. Kun *et al.* in the eleventh article of this issue evaluate financial documents, in which same words may convey different sentiments across different sectors. Therefore, if documents from multiple sectors are learned simultaneously, performance of any learning system can deteriorate. The authors conducted sentiment analysis of 8000 financial reports of firms, sector by sector. Their experiment show that their proposed approach improves prediction performance by substantially over the baseline model, and that the direction of post-announcement stock price movements shifts accordingly with the polarity of the sentiment of reports. And finally, Wang *et al.* in the last article of this issue argue that privacy protection should be carried out first because it contains personal sensitive information. To reduce the computational complexity and low speed of existing privacy-preserving algorithms for high-dimensional data publishing, the authors propose a probabilistic optimal projection partition k-dimensional (KD)-tree k-anonymity algorithm. The proposed algorithm is validated by a theoretical analysis and comparison experiments. The results show that the proposed algorithm can reduce the average generalization range compared to traditional k-anonymity.

In conclusion, we would like to thank all the authors who have submitted the results of their excellent research to be evaluated by our referees and published in the IDA journal, over the last 22 years. We

look forward to receiving your feedback along with more and more quality articles in both applied and theoretical research related to the field of IDA.

With our best wishes,

Dr. A. Famili
Editor-in-Chief