

## Editorial

---

Dear Colleague:

Welcome to volume 16(6) of Intelligent Data Analysis Journal

This issue of the IDA journal consists of nine articles which represent a variety of topics, all related to the applied and theoretical research in the field of Intelligent Data Analysis.

The first three articles are about data understanding and data preprocessing. Abellán and Moral in the first article discuss the issue of attribute dependency and introduce a method to identify when two categorical variables are independent. They compare their method with different scores such as Bayesian metric, Bayesian information criterion, the p-value of the Chi-square test, where the results are shown to be more suitable for certain type of problems. Nikolić's article is about measuring similarity of graph nodes where a method is proposed to perform sorting based on refined concept of node similarity through matching neighbors. The experimental results show the convergence of this approach where it is evaluated on several problems and its comparison to other methods. Makarehchi and Kamel in the next article of this issue discuss the lack of robustness in feature ranking of text when it is used in different data sets. They introduce a new method that is based on combining feature rankings and selection of the best features. They apply their proposed method in text classification problems and evaluate it on three well-known data sets using support vector machines and Rocchio classifier. They show that combining methods can offer reliable results.

The next two articles are mostly on classification. Li *et al.* in the fourth article argue that SVM's that are highly accurate classifiers, are not easy to use in large data sets because of their training complexity. They propose a new approach to SVM that relies on a two stage process that consists of a random selection to select a small group of training data followed by a de-clustering technique to recover training data for a second stage SVM. From their experimental results on several data sets the distinctive advantage of their approach on analyzing large data sets is shown. Sameon *et al.* in the next article argue that conventional threshold selection in Boolean reasoning based discretization produces unacceptable results and propose a solution that includes particle swarm optimization. They argue that the first task is to reduce the search space and the second task is to reconstruct the fitness function. Their results of experimentation using four real data sets show how their approach outperforms the existing discretizers in terms of classification accuracy as well as reduction of the decision rules.

Frequent item sets is the topic of the next two articles. Rajalaxmi and Natarajan in the sixth article of this issue discuss privacy preserving data mining and propose an approach that consists of hiding sensitive utility and frequent itemsets and it contains minimum impact on the non-sensitive information. Results of their evaluation on synthetic and real data sets show the effectiveness of their approach in minimizing the non-sensitive itemsets as well as maintaining data quality in the sanitized database. Nguyen and Yamamoto in the next article propose an incremental mining algorithm for closed frequent labeled ordered trees. This is based on the adaptation of a divide-and-conquer strategy where different data mining strategies are applied at different stages of the data mining process. Their experiments demonstrate the efficiency and scalability of their algorithm on both synthetic and real data sets.

The last two articles are more along the line of benchmarking and applied research. Sun and Garibaldi argue that parametric and non-parametric approaches are suitable for model selection and discuss robust

clustering where they use student t-distribution as the building block. They introduce two clustering algorithms where one is based on student t-distribution and the other with Dirichlet Process mixture. They compare the new algorithms on a set of UCI data sets where their experimental results show the advantages of their proposed methods on identifying outliers. Rasheed and Alhadjj in the last article of this issue discuss the importance of finding periodic patterns in large data sets and analyze a set of stock market data in detail. They demonstrate how the periodic pattern analysis may help in better understanding of the data to predict the future trends. Their experimental results show that taking into account missing values may result in identifying more interesting patterns in the data.

In conclusion, with the completion of Volume 16 of the IDA journal, we continue receiving more and more quality papers submitted to the IDA journal for evaluation and publication. Over the next year, we will have 1–2 special issues to be produced from the best papers of related conferences and workshops. The IDA conference that is usually organized in Europe, was would be held in Helsinki, Finland from October 25–27, 2012. We will provide some details in the next issue of the IDA journal. 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*