

Editorial

Dear Colleague:

Welcome to volume 15(4) of Intelligent Data Analysis Journal.

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

The first three articles are more aligned with unsupervised learning. In the first article of this issue, Yanto et al. investigate a new approach to data clustering in which they propose a data clustering approach using rough set theory to deal with problems in categorical data. The algorithm is based on a variable precision rough set model. Their results show that the proposed approach provides better performance in selecting the clustering attributes. Shani et al. in the second article focus on unsupervised segmentation of sequence of symbols into chunks and present an unsupervised learning algorithm to perform this task. Their algorithm that builds a lexicon of segments and computes a maximum likelihood segmentation given the current lexicon, is most appropriate for hierarchical sequences where smaller segments are grouped into larger segments. This algorithm is compared to previous approaches and the authors present segmentation results from the interactions of a web service and its clients. Hajian et al. in the third article of this issue discuss the problem of data privacy in data mining and introduce a privacy preserving clustering technique suitable for horizontally and vertically distributed data sets. Their proposed method contains some form of wavelet transforms and scaling data perturbation to achieve both data hiding and data reduction so that all numeric attribute values are protected. The experiments presented in this article have shown that the proposed technique provides better privacy and clustering results comparing to existing methods.

The next four articles of this issue are on pattern discovery. Umek and Zupan emphasize the need for identifying subsets of data instances with similar outcome description and present a method that directly addresses this problem. Their experiments with the proposed method on a set of synthetic and real life social survey data demonstrate the strength of their approach to discover relevant and interesting subgroups from the data with multi-dimensional purposes. In the fifth article of this issue, Yun and Ryu discuss different aspects of the topic of weight based sequential pattern mining and emphasize that in noisy environments a small change in weights or supports of items affects the results. They propose a robust concept of mining weighted approximate sequential patterns in which an approximate factor is defined to relax the requirement for exact quality between weighted supports of sequential patterns and a minimum threshold. Their proposed approach is extensively evaluated in several experiments and the results are presented in the paper. Deypir and Sadreddini in the next article discuss some of the challenges in mining frequent patterns such as reduction in runtime and memory usage and propose a novel method for efficient mining of frequent patterns in data streams. Their proposed method is based on a sliding window model which divides selected windows into a number of panes. This approach that dynamically adapts itself with concept change is empirically evaluated against some existing methods. Their experimental results on synthetic and real life data streams show the better performance of their approach on runtime and memory usage.

Zliobaite's article on concept drift explains the challenge in supervised learning for sequential data. It includes a method for training set selection which is more suitable for cases where concept drift is gradual. The proposed method dynamically determines an optimal training set size at every time step using a cross validation approach. The experimental results show the best accuracy in the peer group on the real and artificial drifting data. Ordonez in the next article of this issue discusses data preprocessing and data transformation in data base systems, which is an important and yet time consuming task in data mining. The article, being also an overview of data preprocessing methods, presents a set of examples and common solutions to data preparation and transformations. It also highlights which steps in data mining become more efficient when proper data preprocessing is performed. And finally in the last article of this issue, Kimura et al. address the problem of parameter estimation in observed data of a social network. To do this, the authors formulate the likelihood to obtain the observed data which is a set of time sequence cases of infected nodes and propose an iterative method to search for the parameters that maximize this likelihood. Their results show the effectiveness of their approach on large data sets related to social networks.

In conclusion, we continue our efforts for organizing and publishing more special issues of IDA journal. We would also like to remind our readers about the IDA conference that will be held in Porto, Portugal from October 29–31, 2011. For details, conference dates and other information please refer to the following web site (<http://www.liaad.up.pt/ida2011/>). 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