

Editorial

Dear Colleague:

Welcome to volume 11(5) of *Intelligent Data Analysis* – An international Journal.

This issue of the IDA journal consists of seven articles that represent a collection of some very interesting applied and theoretical research in Intelligent Data Analysis. The first four articles are on different areas of learning while the last three are primarily on data pre-processing.

The first article by Hasler and Hornik discuss the problem of identifying the best measure of interestingness in discovering association rules and present a simple probabilistic framework that can be used to simulate transaction data when no associations are present. Their result show that using their approach, confidence is systematically influenced by the frequency of the items in the antecedent part of an association rule. The evaluation of two interest measures introduced in this research show that the new measures significantly perform better than existing ones such as lift, for applications where spurious rules are problematic. The second article by Horman and Kaminka is about unsupervised sequence learning where current approaches are frequency and statistical based. The authors provide an in-depth empirical comparison and analysis of popular sequence learning methods in terms of quality of information produced. Focusing on the limitations of the frequency and statistical based approaches, they develop a method that involves normalizing rankings of candidate patterns. Their results show significant improvements in the quality of the results in all methods and across many noise settings. Focusing on developing user models, Srinivasa *et al.* in the next article, propose a combination of Adaptive Genetic Algorithms and multi-class Support Vector Machines to learn user models in the Internet domain. The application is to adapt to the user interests based on XML documents. Their evaluation, which is based on a wide range of documents, indicate that their proposed approach significantly improves the performance of the selective dissemination task with respect to the accuracy and efficiency.

Seewald, in the fourth article of this issue, describe an in-depth analysis of spam-filtering of a proposed Naïve-Bayes learner and two extended variants of it. The three learners were evaluated using 65,000 e-mails from seven users along with 25,000 e-mails from a single user. With the objective of evaluating the performance of the two variants of Naïve-Bayes, the author found that the simpler Naïve-Bayes learner offered the most stable performance and it deteriorated least over time.

A large part of learning requires proper handling of the data, using methods such as boosting, bagging and developing wrappers. Flores *et al.* in the next article propose a supervised wrapper approach to discretization. Their approach involves simultaneous discretization of all variables where they evaluate their approach using a Naïve-Bayes classifier. Their search for optimal discretization is based on a set of separate evolutionary algorithms. They perform an analysis of different parameters of their algorithm using ANOVA and a number of artificial as well as some UCI data sets. Their results suggest that their method is more effective and robust for variable discretization.

Along the same line data preprocessing, Karmaker and Kwek in the next article, propose a data cleaning approach that smoothes out a substantial amount of attribute noise and handles missing attribute values as well. Based on Expectation Maximization, their approach iteratively refines each attribute value of a data set using a predictor constructed from the previously refined values. Their evaluation

includes improved classification accuracy on a number of real data sets from UCI and show that their technique can be easily adapted to fill up missing attribute values in classification problems. The last article of this issue, by Salvador and Chan, introduce a fast procedure for dynamic time wrapping that is linear in time and space complexity. Their approach consists of several levels that recursively project a solution, starting with coarser resolution until a finer one is obtained. Their results include theoretical and empirical evaluation of their approach and its analysis for the accuracy and comparison with two existing algorithms.

And finally, the IDA-2007 (<http://www.ida2007.org/>) conference was held in September in Slovenia this year. We are planning to work with authors of the best 5–6 papers presented during this event and publish them in a special issue of the IDA journal, sometimes in 2008.

With our best wishes,

Dr. A. Famili
Editor-in-Chief