DOI 10.3233/IDA-2010-0467
IOS Press

Editorial

Dear Colleague:

Welcome to volume 15(3) 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.

In the first article of this issue, Dos Santos *et al.* propose an approach that is based on the fundamental principles of heuristic search Bayesian classifiers. The authors use a combination of Markov Blanket concept as well as a newly proposed approximate Markov Blanket to reduce the number of nodes that normally form a Bayesian network. This approach could result in reducing the high computation cost of heuristic search in learning algorithms. Their approach is evaluated using data from twelve domains and is compared with Naïve Bayes and Tree Augmented Network Classifiers. The second article of this issue by Abellán *et al.* is also about Bayesian classifiers in which they present a semi-Naïve Bayes classifier that searches for dependent attributes using a filter approach. Their method uses a grouping procedure following each merge of variables and groups two or more cases of the new variable into a single one. Their proposed approach is a competitive classifier based on class probability estimates.

Qu *et al.* in the third article of this issue, propose to modify support vector machines (SVM) to make it more suitable for highly imbalanced overlapping classification cases. The main motivation for this work was that in SVM based algorithms the increase in the number of correctly predicted minority samples will lead to more majority samples to be misclassified. Their proposed approach can identify non-overlapping samples in one feature space and shifting kernel spaces in different spaces. Their article includes a number of case studies that show the effectiveness of their approach. Krstnic and Slapnicar in the fourth article propose a mode seeking and clustering procedure that is based on the estimation of the discrete probability density function of the data. This density function is estimated in two steps that consist of counting data samples and assignment of bandwidth kernel to each cell. From the results presented in the article, it seems that the proposed technique does not require any assumptions about the structure of the data and is highly efficient.

In the fifth article Thang *et al.* address the robustness issue of maximum likelihood based methods in data clustering where the algorithms are usually sensitive to noise and outliers in the data. They introduce a variant of classical mixture model based clustering that includes a genetic algorithm component. Their analytical and experimental studies show that their approach can overcome the negative impact of outliers in data clustering which results in identifying more accurate and reliable clustering results. Mo and Huang in the next article discuss the importance of proper feature selection in data preprocessing stage of data analysis and introduce a unified dependency criterion called inference correlation. The approach is evaluated as part of a feature selection algorithm on a number of synthetic and real data sets where the results confirm the effectiveness of their approach. The next article by Yang *et al.* discusses evidence conflict in information fusion and emphasize on the stochastic interpretation for basic probability assignment. They compare several approaches to deal with evidence conflict and propose a new algorithm that is a combination of absolute and relative difference factors of two pieces of evidence.

Their algorithm is evaluated using a numerical example where the analysis shows the efficiency of their approach.

Data Mining in software in engineering is the subject of the next article by Halkidi *et al.* where they describe various data sources and discuss the principles and techniques of data mining applied to software engineering data. This article surveys several data mining approaches used in software engineering and groups them according to the corresponding parts of the software development that they are appropriate for. This grouping is useful for selection of appropriate techniques. An finally, Gonzalez *et al.* in the last article of this issue present a method to identify leukemia patients from bone marrow cell images using a combined machine vision and data mining strategy. The authors show how a combination of descriptive features and Eigenvalues help to improve classification accuracy which exceeds ranges of 90–95% in the data used in their studies.

In conclusion, we are planning to have two special issues this year, one based on a proposal and another one based on a workshop held during one of the related conferences. We would like to remind the readers about the IDA conference that will be held in Porto-Portugal from October 29–31, 2011. For details, important 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.

With our best wishes,
Dr. A. Famili
Editor-in-Chief