Intelligent Data Analysis 20 (2016) 1243–1245
DOI 10.3233/IDA-160865
IOS Press

Editorial

Dear Colleague:

Welcome to volume 20(6) of Intelligent Data Analysis (IDA) Journal.

This issue of the IDA journal, the sixth and the last issue of 2016, consists of twelve articles, all covering a wide range of topics related to the theoretical and applied research in the field of Intelligent Data Analysis.

The first two articles of this issue are about various aspects of data preprocessing and outlier detection in IDA. Yu *et al.* in the first article discuss the topic of outlier detection and propose an information-entropy-based k-nearest neighborhood outlier detection algorithm that is combined with Shannon information theory and the triangle pruning strategy. The algorithm accounts for the data points whose k-nearest neighbors are distributed on the edge of the range within the designated radius. Their experimental results show that, compared to existing methods, their proposed approach improves pruning and detection rates while maintaining the coverage rate. Han *et al.* in the next article discuss the issue of incomplete information systems with missing or unknown data that may affect the quality of data-driven decision fusion. The authors recommend the rough set theory for the decision fusion of incomplete information systems and propose a new approach to evaluate the impact of missing data. They define an α-classification quality of approximation to measure the quality of decision fusion with various identical degrees. Their experimental results show that the quantitative evaluation of missing data in an existing information system can be made by the proposed method and the volume of acceptable missing data according to a determined quality is possible to be predicted in future applications.

The next two articles are on the topic of support vector learning. Bhavsar and Ganatra in the third article of this issue present a classification algorithm that uses the SVM in the training phase and the Mahalanobis distance in the testing phase, in order to design a classifier which has low impact of kernel function on the classification accuracy. They use the Mahalanobis distance to replace the optimal separating hyper-plane as the classification decision making function in the SVM. Their experimental results show that the accuracy of their proposed SVM classifier has a low impact on the implementation of kernel functions. Zhao *et al.* in the fourth article of this issue propose a novel method called Lap-NPSVM for binary classification under a Semi Supervised Learning scenario. One of the main merits of this approach is that it avoids inversion matrix in objective function compared with Laplacian SVM and other twin SVM based approaches which is obviously a big obstacle for large scale applications. Their experiments on artificial and real world datasets show the generalization and speed effectiveness of this approach.

The third group of articles in this issue are on data streams Dehghan *et al.* discuss the topic of concept drift in data streams. They emphasize that an ideal method for concept drift detection should be able to rapidly and correctly identify changes in the underlying distribution of data points and adapt its model as quickly as possible while the memory and processing time is limited. The authors propose a novel method based on ensemble classifiers for detecting concept drift. The proposed approach processes samples one by one, and monitors the distribution of ensemble's error in order to detect probable drifts. Their experimental results show that the proposed method is capable of detecting and adjusting to concept drifts from different types, and it has outperformed well-known state-of-the-art methods, especially, in the case of high-speed concept drifts. Safaei *et al.* in the next article argue that data stream processing has real-time requirements and due to the memory limitations of Data Stream Management System, we normally need one-pass algorithms. The authors employ a method for answering Ad-hoc Continuous Aggregate Queries over data streams that uses a Dynamic Prefix Aggregate Tree. As each tuple from data stream arrives, the required data is stored as a tree structure and used when applying the aggregate functions. It has been found out both empirically and analytically that the proposed method is more cost-effective than using the conventional Prefix Aggregate Trees, in terms of time and memory capacity.

The last group of articles in this issue are on novel applications of IDA methods. Louveaux *et al.* in the first article of this group argue that many industrial companies monitor their production process, collecting large amount of measurements. The authors describe a technique using this data to improve the performance of a monitored process. In particular their goal is to find a set of rules, i.e. intervals on a reduced number of parameters, for which an output value is maximized. This article compares a machine learning-based heuristic to the solution computed by a mixed-integer linear program on real-life databases from steel and glass manufacturing. Computational results presented in this article show that the heuristic obtains comparable solutions to the mixed integer linear approach. Li *et al.* in the eighth article of this issue argue that conventional dictionary learning algorithms mainly focus on reconstructing the training samples and cannot directly associate the learning procedure with the test samples. The authors present a test sample oriented two-phase dictionary learning (TSOTP-DL) algorithm suitable for face recognition where in the first phase all training samples are used to provide a linear representation of the test sample, and in the second phase a dictionary is learned for the test sample by using the selected “important” training samples. Their experiment results demonstrate that the proposed algorithm achieves better classification results than some state-of-the-art dictionary learning and sparse coding algorithms on four public face databases. In the ninth article of this issue Vanitha *et al.* present a multi-SVM approach with a novel gene selection method using Mutual Information (MI) that is suitable for multi-class classification in the cancer diagnosis. The proposed approach constructs separate classifiers for each class and the combined multi-class classifier assigns a cancer tissue sample to the class with the highest support. The performance of their proposed Multiclass Support Vector Machine (mSVM) with Gene Selection using the mutual information approach is evaluated on four benchmark gene expression datasets for cancer diagnosis, From the simulation study presented, it is observed that the proposed approach reduces the dimensions of the input features by identifying the most discriminating gene subset for each category and improves the predictive accuracy for multi-class cancer. The next article by Tsai and Lu is on applying Markov-Exponential Grey Model (EGM) for Forecasting Management. The authors present a mechanism that integrates a grey model with an exponentially weighted moving average adjustor, known as the EGM. This novel EGM model is combined with Markov processes to generate a precise models for forecasts. Their experimental results show that the average precision of their models improves substantially. Zliobaite and Khokhlov in the eleventh article of this issue discuss mobile route planning based on data collected from GPS technology in urban environments and argue that one of the main challenges for travel time estimation and prediction in such a setting is how to aggregate data from vehicles that have followed different routes, and predict travel time for new routes of interest. It seems that the optimal way to estimate travel times to minimize the expected mean absolute error is to combine the mean and the median times on individual segments, where the combination function depends on the number of segments in the route of interest. The authors present a methodology for obtaining such estimates, and demonstrate its performance on a case study using travel time data from a district of St. Petersburg collected over one year. And finally, Argueta *et al.* in the last article of this issue argue that if microblogs are analyzed and interpreted correctly they can provide useful information such as understanding how people feel or react towards a specific topic. The authors propose an unsupervised graph-based algorithm to extract emotion bearing patterns from micro-blog posts. Having the extracted patterns, a classifier is implemented to efficiently identify the emotions expressed in posts without depending on predefined emotional dictionaries, lexicons or ontologies. Experimental results are shown for English, Spanish and French tweets where they achieve a desired accuracy, generality, adaptability and minimal supervision.

In conclusion, with the completion of this issue of the IDA journal, we would like to let you know that the IOS Press, the publisher of the IDA journal, is celebrating the 20th anniversary of founding the IDA journal. This year, in addition to our six regular issues, we also published a special issue related to CIARP series of conferences. 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