Intelligent Data Analysis 19 (2015) 1195–1197
DOI 10.3233/IDA-150777
IOS Press

Editorial

Dear Colleague:

Welcome to volume 19(6) of Intelligent Data Analysis (IDA) Journal.

This issue of the IDA journal, the sixth and last issue of 2015, 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 four articles are on various forms of data preprocessing. He *et al.* in the first article of this issue argue that label noise will create all kinds of problems in the feature selection procedure and the performance of any classifier. Using Parzen window they propose a novel mutual information estimator which is based on a probabilistic label noise model. Their experiments over a toy dataset and eight real world datasets, while performing classification with a kNN classifier, they demonstrate that their proposed approach is sound and able to reduce the influence of label noise effectively. Strange and Zwiggelaar discuss that reducing the dimensionality of a dataset whilst retaining its inherent manifold structure is key to many tasks in pattern recognition and machine learning. The authors propose a heuristic approach to tackle this problem that involves approximating the manifold as a set of piecewise linear models. A detailed analysis of the proposed approach is presented along with comparison with existing manifold learning techniques. Results presented in their paper on both artificial and image based data show that in many cases this heuristic approach to manifold learning is able to out-perform traditional techniques. In the third article of this group Mohammadi *et al.* argue that analysis of network traffic, financial transactions, and mobile communications are examples of applications where examining entire samples of a large dataset is computationally expensive, and requires significant memory space. To reduce the number of samples without compromising the accuracy, the authors propose a new cluster-based sample reduction method which is unsupervised, geometric, and density-based. The performance of the proposed method is measured on various datasets and compared with several cluster-based and density-based methods. They demonstrate that, while reducing the sample size of the input dataset in half, the classification accuracy is not reduced significantly, indicating that the proposed method selects the most relevant samples from the original dataset. In the last article of this group on data preprocessing, Maldonado and Lopez address the issue of high dimensionality of feature selection for linear and kernel-based Support Vector Machines and propose an embedded feature selection approach for Support Vector classification. This approach is based on a sequential backward elimination which uses different linear and kernel-based contribution measures to determine the feature relevance. Their experimental results with microarray datasets demonstrate the effectiveness in terms of predictive performance and construction of a low-dimensional data representation.

The next group of articles are about classification and unsupervised learning. Amorim and Cardoso in the first article of this group compare clustering solutions using indices of paired agreement and propose a new method called IADJUST to correct indices of paired agreements, excluding agreements by chance. The authors illustrate its use in external clustering validation, to measure the accordance between clusters and an *a priori* known structure. Their experiments involve simulated data sets, under a range of scenarios – considering diverse numbers of clusters, clusters overlaps and balances where they discuss

the pertinence and the precision of their approach. Gutierrez-Rodriguez in the next article introduce an algorithm for extracting a small subset of patterns useful for clustering. Their proposed algorithm extracts patterns from a collection of trees generated through a new induction procedure. Their experimental results show that the proposed algorithm extracts significantly less patterns in a relatively less time than recent pattern-based clustering algorithms, but obtaining similar clustering results in terms of F-measure. In addition, this proposed algorithm obtains similar clustering quality results than traditional clustering algorithms. In the seventh article of this issue, Mirzaie *et al.* argue that producing overlapping clusters is a major issue in clustering methods and explain that most of the research in this area has focused on clustering using disjoint clusters. This becomes a major problem in some domains such as microarray datasets where many gene regulatory networks have inherently overlapping partitions. The authors propose a new density based clustering which includes a bound on the number of overlapping clusters and a closeness measure. The authors compare their proposed approach with DBscan (a non-overlapping density-based clustering) algorithm and prove that their approach could be significantly better than non-overlapping clustering in microarray data. Ebrahimi-Dishabi and Abdollahi-Azgomi in the last article of this group focus on the topic of privacy preserving clustering where the objective is to preserve the privacy of data during clustering analysis. They propose a differential-based algorithm that is suitable for horizontally and vertically distributed datasets. The authors use orthogonally discrete wavelet transforms (DWT) to obtain perturbed data with both low data dimensionality and less noise addition. Their experiments involve analyzing some well-known datasets where the results show that their proposed algorithm guarantee an appropriate level of both utility and privacy of the published data.

The last four articles of this issue are on learning and prediction. Lofstrom *et al.* discuss the topic of label-conditional conformal prediction, which is a specialization of the framework which gives a bound on the error rate for each individual class. This would be quite useful for imbalanced data sets where many learning algorithms have a tendency to predict the majority class more often than the expected relative frequency, i.e., they are biased in favor of the majority class. The study reports the author's investigation into the class bias of standard and label-conditional conformal predictors and the analysis of a number of publicly available datasets with varying degrees of class imbalance. Their experimental results show that while conformal prediction is highly biased towards the majority class on imbalanced datasets, label-conditional conformal prediction is not biased towards the majority class. Hussein *et al.* in the next article of this group argue that the size and the unstructured nature of the information makes the location of the right information a challenging task in internet. Recommender systems and web usage mining techniques are two of the main methods that are apparently used. The authors present a framework for the next page prediction that exploits users' access history combined with his semantic interests to generate personalized and accurate recommendations. The suggested framework employs user clustering to focus the search which substantially reduces prediction time. In the eleventh article of this issue, Silva and Cardoso discuss generalized additive models and argue that scorecards, the discretized version of these models, are a long-established method due to its balance between simplicity and performance. The authors address scorecard development, introduce a new formulation that is more suitable to support regularization and tackle both the binary and the ordinal data classification problems where their proposed methodology shows advantages when evaluated using real datasets. And finally, Su *et al.* in the last article of this issue discuss the skew sensitivity of random forest and rotation forest due to use of Gini index and information gain. In order to improve the performance of these two learning methods, the authors propose using Hellinger distance as the splitting criterion for building each tree in random forest and rotation forest. Their experimental results demonstrate that using Hellinger distance as the splitting criterion to build individual decision tree can improve the performances of both methods, especially for highly imbalanced classifications.

In conclusion, with this issue of the IDA journal, which is Volume 19(6), we are gradually celebrating the 20^{th} anniversary of our journal. The IOS press office, the publisher of the IDA journal, has several plans for this year and mostly for next year. In addition to our six regular issues that now contain 11–12 articles, our plan starting last year has been to publish one special issue per year, which is normally related to a scientific conference for which organizers have submitted an interesting proposal. 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