Neural Comput & Applic (2016) 27:2153–2155
DOI 10.1007/s00521-016-2327-3

EDITORIAL

# Guest editorial: special issue on predictive analytics using machine learning

**Ali Bou Nassif[1] · Mohammad Azzeh[2] · Shadi Banitaan[3] · Daniel Neagu[4]**

Published online: 23 April 2016


Predictive analytics is concerned with the prediction of future trends and outcomes. The approaches used to conduct predictive analytics can be classified into machine learning techniques and regression techniques. Machine learning techniques have become increasingly popular in conducting predictive analytics due to their outstanding performance in handling large-scale datasets with uniform characteristics and noisy data. Innovative predictive models have been applied successfully in several domains such as health care, cyber security, education, credit card fraud detection, social media, cloud computing, software measurement, quality and defect prediction, cost and effort estimation and software reuse.

The main focus of this special issue is on the current state of practice of machine learning techniques to address various predictive problems. A total of 14 papers are included in this special issue. A summary of the papers appears below.

In "Dynamic neural networks for gas turbine engine degradation prediction, health monitoring and prognosis" by Kiakojoori and Khorasani, the problem of health monitoring and prognosis of aircraft gas turbine engines using intelligent-based methodologies are investigated to support enhancement of flight safety and performance. Two different neural network-based architectures, namely the nonlinear autoregressive neural network with exogenous input (NARX) and the Elman neural network, are proposed and developed for predicting the health condition of the engine. Various degradations may occur in a given engine resulting in changes in the performance of its components. Two main degradation factors, namely the compressor fouling and the turbine erosion, are modeled and considered in this work. The proposed dynamic neural networks are first developed and applied to capture the degradation dynamics in the aircraft jet engine. The health condition of the engine is then predicted subject to occurrence of these deteriorations. Multiple scenarios are considered and extensive simulations are conducted corresponding to both dynamic neural network approaches. For each degradation scenario, several neural networks are trained and their performance in predicting the turbine exhaust temperature for multi-flights ahead is evaluated. The most suitable neural network for this prediction is then selected by using a normalized Bayesian information criterion model selection. Simulation results presented demonstrate and illustrate the effective performance of the proposed dynamic neural network-based prediction and prognosis strategies.

In "A hybrid grid-GA-based LSSVR learning paradigm for crude oil price forecasting", Yu et al. propose a hybrid learning paradigm integrating least squares support vector regression (LSSVR) with a hybrid optimization searching approach for the parameters selection in the LSSVR in order to effectively model crude oil spot price with inherently high complexity. The grid method is first applied to determine the proper boundaries of the parameters in the LSSVR. Then, the genetic algorithm is implemented to

✉ Ali Bou Nassif
anassif@sharjah.ac.ae

[1] Department of Electrical and Computer Engineering, University of Sharjah, Sharjah, UAE

[2] Department of Computer Software Engineering, Applied Science University, Amman, Jordan

[3] Department of Mathematics, Computer Science, and Software Engineering, University of Detroit Mercy, Detroit, MI, USA

[4] School of Electrical Engineering and Computer Science, University of Bradford, Bradford, UK

select the most suitable parameters. For illustration and verification, the proposed learning paradigm is used to predict the crude oil spot prices of the West Texas Intermediate and the Brent markets. The empirical results demonstrate that the proposed hybrid paradigm can outperform its benchmarking models in terms of both prediction accuracy and time-savings.

The main objective of image parsing is to understand and perceive the actual content of a given image. In "A crowding multi-objective genetic algorithm for image parsing", Joseph and Auwatanamongkol propose a crowding genetic algorithm for image parsing. The experimental results show that the proposed algorithm performs better than the existing methodologies, in terms of class-wise and pixel-wise accuracy. The results also show that the proposed method can effectively identify multiple object instances existing in a given image.

In "Metaheuristic optimization of multivariate adaptive regression splines for predicting the schedule of software projects", Ferreira-Santiago et al. propose a new model that is derived from the multivariate adaptive regression splines (MARS) model. The new model, optimized MARS (OMARS), uses a simulated annealing process to find a transformation of the input data space prior to applying MARS in order to improve accuracy when predicting the schedule of software projects. The prediction accuracy of the OMARS model is compared to the stand-alone MARS and a multiple linear regression (MLR) model with a logarithmic transformation. The two independent variables used for training and testing the models are functional size and the maximum size of the team of developers. The dataset of projects is obtained from the International Software Benchmarking Standards Group (ISBSG) Release 11. Results based on the absolute residuals and t paired and Wilcoxon statistical tests show that prediction accuracy with OMARS is statistically better than that with the MARS and MLR models.

The main issue when using analogy-based effort (ABE) estimation is how to adapt the effort of the retrieved nearest neighbors. In "Pareto efficient multi-objective optimization for local tuning of analogy-based estimation", Azzeh et al. show that there are three interrelated decision variables that have great impact on the success of adaptation method: (1) number of nearest analogies ($k$), (2) optimum feature set needed for adaptation and (3) adaptation weights. The main theme of the proposed approach is how to come up with best decision variables that improve adaptation strategy and thus the overall evaluation measures without degrading the others. The authors propose to view the building of adaptation procedure as a multi-objective optimization problem. The particle swarm optimization (PSO) algorithm is utilized to find the optimum solutions for such decision variables based on optimizing multiple evaluation measures. The experimental results show that: (1) predictive performance of ABE has noticeably been improved, (2) optimizing all decision variables together is more efficient than ignoring any one of them, and (3) optimizing decision variables for each project individually yields better accuracy than optimizing them for the whole dataset.

Many existing recommendation methods such as matrix factorization (MF) mainly rely on user–item rating matrix, which sometimes is not informative enough, often suffering from the cold-start problem. In "Two-level matrix factorization for recommender systems", Li et al. incorporate the complementary textual relations between items into recommender systems (RS) to solve the aforementioned problem. The authors first apply a novel weighted textual matrix factorization (WTMF) approach to compute the semantic similarities between items, then integrate the inferred item semantic relations into MF and propose a two-level matrix factorization (TLMF) model for RS. Experimental results on two open datasets not only demonstrate the superiority of TLMF model over benchmark methods, but also show the effectiveness of TLMF for solving the cold-start problem.

Class decomposition describes the process of segmenting each class into a number of homogeneous subclasses. Utilizing class decomposition can provide a number of benefits to supervised learning, especially ensembles. It can be a computationally efficient way to provide a linearly separable dataset without the need for feature engineering required by techniques like support vector machines. For ensembles, the decomposition is a natural way to increase diversity, a key factor for the success of ensemble classifiers. In "A fine-grained random forests using class decomposition: an application to medical diagnosis", Elyan and Gaber propose to adopt class decomposition to the state-of-the-art ensemble learning random forests. Medical data for patient diagnosis may greatly benefit from this technique, as the same disease can have a diverse of symptoms. The authors have experimentally validated their proposed method on a number of datasets that are related to the medical domain. Results show that the proposed method has significantly improved the accuracy of random forests.

In "Classifying component failures of a hybrid electric vehicle fleet based on load spectrum data", Bergmeir et al. propose a parameter tuning framework that enables the random forest models, formed by univariate and multivariate decision trees, respectively, to handle the class imbalance problem of the dataset and to select only a small number of relevant variables in order to improve classification performance and to identify failure-related variables. By achieving an average balanced accuracy value of 85.2 %, while reducing the number of variables used from 590 to 22 variables, the results for failures of the hybrid car

battery (approximately 200 faulty, 7000 non-faulty vehicles) demonstrate that especially balanced random forests using univariate decision trees achieve promising classification results on load spectrum data.

In "Deep multilayer multiple kernel learning", Rebai et al. proposed a backpropagation multilayer of multiple kernel learning (MLMKL) framework. Specifically, they propose to optimize the network over an adaptive backpropagation algorithm. The authors use the gradient ascent method instead of dual objective function, or the estimation of the leave-one-out error. The authors test their proposed method through a large set of experiments on a variety of benchmark datasets. Empirical results over an extensive set of experiments show that the proposed algorithm achieves high performance compared to the traditional MKL approach and existing MLMKL methods.

In "Supervised classification of spam emails with natural language stylometry", Shams and Mercer report the development and evaluation of an anti-spam filter based on natural language and stylometry attributes. The proposed filter extracts natural language attributes from email text that are closely related to writer stylometry and generate classifiers using multiple learning algorithms. Experimental outcomes show that classifiers generated by meta-learning algorithms such as ADABOOSTM1 and BAGGING are the best, performing equally well and surpassing the performance of a number of filters proposed in previous studies, while a random forest generated classifier is a close second. On the other hand, the performance of classifiers using support vector machine and Naive Bayes is not satisfactory.

In "A Q-learning-based swarm optimization algorithm for economic dispatch problem", Hsieh and Su treat optimization problems as a kind of reinforcement learning problems regarding an optimization procedure for searching an optimal solution as a reinforcement learning procedure for finding the best policy to maximize the expected rewards. This viewpoint motivated the authors to propose a Q-learning-based swarm optimization (QSO) algorithm. The proposed QSO algorithm is a population-based optimization algorithm which integrates the essential properties of Q-learning and particle swarm optimization. The optimization procedure of the QSO algorithm proceeds as each individual imitates the behavior of the global best one in the swarm. The best individual is chosen based on its accumulated performance instead of its momentary performance at each evaluation. Two datasets including a set of benchmark functions and a real-world problem—the economic dispatch (ED) problem for power systems—are used to test the performance of the proposed QSO algorithm. The simulation results on the benchmark functions show that the proposed QSO algorithm is comparable to or even outperforms several existing optimization algorithms. As for the ED problem, the proposed QSO algorithm has found solutions better than all previously found solutions.

In "Offer acceptance prediction of academic placement", Shrestha et al. examine the validation of prediction models of acceptance of academic placement offers by students in the context of international applications at a large metropolitan Australian University. The important predictors for the acceptance of offers are as follows: the chosen course and faculty, whether the student was awarded any form of scholarship and also the visa assessment level of the country by the immigration department. Prediction models are developed using a number of classification methods such as logistic regression, Naive Bayes, decision trees, support vector machines, random forests, k-nearest neighbor, neural networks, and their performances are compared.

In "Neural network models for software development effort estimation: a comparative study", Nassif et al. compare four different neural network models, namely multilayer perceptron, general regression neural network, radial basis function neural network and cascade correlation neural network. The comparisons are based on: (1) predictive accuracy centered on the mean absolute error criterion, (2) whether such a model tends to overestimate or underestimate and (3) how each model classifies the importance of its inputs. Industrial datasets from the International Software Benchmarking Standards Group (ISBSG) are used to train and validate the four models. Results show that the four models tend to overestimate in 80 % of the datasets, and the significance of the model inputs varies based on the selected model. Furthermore, the cascade correlation neural network outperforms the other three models in the majority of the datasets constructed on the mean absolute residual criterion.

In "Combining time series prediction models using genetic algorithm to autoscaling Web applications hosted in the cloud infrastructure", Messias et al. propose an adaptive prediction method using genetic algorithms to combine time series forecasting models. The proposed method does not require a previous phase of training, because it constantly adapts the extent to which the data are coming. To evaluate our proposal, three logs extracted from real Web servers are used. The results show that the proposed approach often brings the best result and is generic enough to adapt to various types of time series.

Finally, the guest editors are grateful to the reviewers for their insightful and timely reviews and to the Editor-in-Chief for giving them the opportunity to propose and guest-edit this special issue.