New Generation Computing, 29(2011)241-244
Ohmsha, Ltd. and Springer

NEW GENERATION COMPUTING

Preface

Special Issue on Hybrid and Ensemble Methods in Machine Learning

Oscar CORDÓN
European Centre for Soft Computing
33600 Mieres, SPAIN
Dept. Computer Science and Artificial Intelligence
E.T.S.I. Informática y Telecomunicación
18071 Granada, SPAIN
oscar.cordon@softcomputing.es
Przemysław KAZIENKO and Bogdan TRAWIŃSKI
Institute of Informatics
Wrocław University of Technology
50-370 Wrocław, POLAND
{kazienko, bogdan.trawinski}@pwr.wroc.pl

Received 31 March 2011

Hybrid and ensemble methods in machine learning have gained great attention of the scientific community over the last few years. Multiple learning models have been theoretically and empirically shown to provide significantly better performance than their single base model counterparts. Ensemble algorithms and hybrid methods of reasoning have found their application in various real word problems ranging from person recognition through medical diagnosis and text classification to financial forecasting.

This special issue encompasses six papers devoted to both hybrid and ensemble methods and their application to classification, prediction, and clustering problems. The issue originated from the special session on "Multiple Model Approach to Machine Learning" (MMAML 2010) organized by the guest editors at the Second Asian Conference on Intelligent Information and Database Systems (ACIIDS 2010), that was held in Hue City, Vietnam, during March 24-26, 2010. Thirteen papers and five posters were selected out of thirty five submissions for presentation at the special session. The authors of the thirteen full papers were invited to submit significantly extended versions of their contributions to the current special issue and eleven submissions were finally received. After a thorough review process, only six of the submitted contributions were finally considered by the guest editors and the journal editors to become part of the issue. Unfortunately, the five remaining contributions were rejected due to the high quality of standards we wanted to impose for the special issue, which resulted in a rejection rate close to fifty percent.

Our aim is that the six accepted contributions succeed at covering the range of the different existing approaches to the wide area of the design of hybrid and ensemble methods for machine learning as much as possible. These six contributions are briefly reviewed as follows.

In the first paper, entitled "Condition-based Maintenance with Multi-Target Classification Models," Mark Last et al. focus on a real-world application area, that of automotive prognosis by predictive maintenance of car subsystems by means of a condition-based maintenance approach. Prognostics is recognized to be a key aspect in condition-based maintenance as it considers the prediction of future faults before those faults really occur. In particular, the contribution introduces a data mining approach to prognosis vehicle failures. The proposed solution involves the use of a multi-target probability estimation algorithm, based on a fuzzy network, which is shown to outperform its single-target counterpart. The experimental study considers a database of sensor measurements and warranty claims with the aim to predict the probability and the timing of a failure in a given subsystem, the car battery.

Chunshien Li and Tai-Wei Chiang are concerned with field of system identification by neuro-fuzzy systems, the most extended hybrid system in the area of computational intelligence/soft computing. In the contribution entitled "Function Approximation with Complex Neuro-Fuzzy System Using Complex Fuzzy Sets — A New Approach," they introduce a new approach to design neuro-fuzzy systems based on the complex fuzzy sets structure. Complex fuzzy sets are extensions of classical fuzzy sets based on novel membership functions defined on a complex-valued state within the unit disc of the complex plane, which provides greater potential space for adaptivity. A hybrid design method with the particle swarm optimization metaheuristic and the classical recursive least squares estimator numerical optimization method is used to identify the set of parameters associated to the complex neuro-fuzzy system model structure. The use of the latter optimization algorithm allows the authors to obtain accurate models in a reduced run time. Comparisons between the novel complex neuro-fuzzy system and different fuzzy rule-based system learning methods, neuro-fuzzy systems, and neural networks in three different applications — two function approximation and a time series forecasting problems — clearly show the performance advantage of the proposed approach.

The remaining four contributions are focused on the design of ensemble methods. Bagging and boosting data resampling approaches have arisen as the most extended approaches to build classifier and model ensembles. The paper, "Classification Performance of Bagging and Boosting Type Ensemble Methods with Small Training Sets" by M. Faisal Zaman and Hideo Hirose takes us a step ahead on this large branch of research by dealing with the obtaining of the most appropriate size for the resampled data sets to learn the individual weak classifiers. The authors empirically analyze the bias and variance contribution to the ensemble classification error by considering different training sample sizes ranging from subsampling ratios of a 10% to a 63% (i.e, the usual bootstraping "magic number") of the original training set size. These more or less reduced training sets selected separately for 20 UCI repository datasets are used to design three different types of ensembles based on three different resampling mechanisms — bagging, boosting (Adaboost), and a hybrid bagging-based methodology called bundling. A full decision tree is considered as the base classifier for the three methods, while two support vector machine-based classifiers, the logistic classifier, and the stabilized linear discriminant classifier are used as the additional bundle of classifiers to be trained on the out-of-bag samples in bundling. The main conclusion drawn from the analysis is that ensembles designed with bundling using small subsamples have significantly lower bias and variance than those obtained from bagging and Adaboost.

In the fourth paper, entitled "Boosting-based Sequential Output Prediction," Tomasz Kajdanowicz and Przemysław Kazienko explore the application of classifier ensembles based on boosting to the sequence prediction problem. In their problem variant, the goal is to classify a label sequence (sequential output) by only taking an independent set of attributes as a base. This classification task shows a high complexity since it involves an extension of multi-label classification, a recent hot topic in the pattern recognition area. In this case, the output space requires the handling of a complex structure (e.g. sequences, trees, or graphs) instead of the usual unstructured set of labels. An adaptation of the classical AdaBoost algorithm to allow it to deal with sequence prediction (AdaBoostSeq) is proposed. An advanced cost function is designed in order to cope with the required sequential output prediction while the J48 decision tree learning method is considered as the base classifier induction technique. The experimental study, developed on five different high-dimensional sequence datasets, revealed the good performance of the proposed approach with respect to some benchmarking methods.

The contribution, "Selection of Heterogeneous Fuzzy Model Ensembles Using Self-adaptive Genetic Algorithms" jointly deals with the two main topics in the current special issue's scope, hybrid machine learning methods and model/classifier ensembles. Magdalena Smętek and Bogdan Trawiński propose the design of heterogeneous model ensembles, that is, ensembles composed of individual models with different structures. The design method is based on the use of bagging for the weak learner (individual model) design and self-adapting genetic algorithms to perform model selection by means of an overproduce-and-choose strategy. The individual models are based on fuzzy model structures learned using 16 different heuristic and genetic learning algorithms over the bootstrap replicates (training data set bags). Three different self-adapting genetic algorithms are proposed to develop classifier selection, aiming to decide a final heterogeneous ensemble composition with a good accuracy-complexity tradeoff. A real-world application related to estate appraisals in Poland is considered to test the performance of the generated ensembles, which showed better predictive accuracy than the homogeneous ensembles built for the same problem.

Finally, in the last contribution to the special issue, Bruno Baruque et al. consider the same area as the previous paper but focusing on classifier ensembles. In their paper, "Hybrid Classification Ensemble Using Topology-preserving Clustering," they propose a novel hybrid intelligent system using both unsupervised and supervised learning to design the ensemble. The learning methodology is based on problem decomposition by considering two different stages. In the first one, the problem space is split into different regions (clusters) according to the data distribution in the input space, using an informed decision provided by a self-organizing map algorithm. The second stage involves the derivation of a simple classifier (naive Bayes and multi-layer perceptron are considered as the supervised base classifiers) to cover each of those regions, thus obtaining a specific weak learner allowing the global ensemble not to suffer from overfitting. When a new pattern is to be classified, the partition method assigns it to the corresponding classifier, which is specialized in the classification of that type of data. Hence, the proposed methodology follows the dynamic strategy in classifier ensemble selection to allow the selection of the most confident classifier to label each test pattern individually. In this way, the overall method is quick, simple, and accurate, as demonstrated in the developed experimental study which compares the performance of different techniques for each of the two methodology stages in seven different UCI repository datasets.

Finally, as guest editors of this special issue, we would like to thank all the authors for their high quality contributions and the referees for their outstanding cooperation, as well as for their interesting comments and suggestions that helped the authors to improve the final versions of their papers. Besides, we sincerely thank Professors Taisuke Sato and Toyoaki Nishida, the Editor-in-chief and the Area Editor (Intelligent Systems) of the New Generation Computing journal, respectively, for providing us with the opportunity to edit this issue and for their close collaboration during the editorial process.