Journal of Universal Computer Science, vol. 19, no. 4 (2013), 457-461
submitted: 3/2/13, accepted: 17/2/13, appeared: 28/4/13 © J.UCS

Hybrid and Ensemble Methods in Machine Learning

J.UCS Special Issue

Przemysław Kazienko
(Wrocław University of Technology, Wrocław, Poland
kazienko@pwr.wroc.pl)

Edwin Lughofer
(Johannes Kepler University, Linz, Austria
Edwin.Lughofer@jku.at)

Bogdan Trawiński
(Wrocław University of Technology, Wrocław, Poland
bogdan.trawinski@pwr.wroc.pl)

Hybrid and ensemble methods in machine learning have attracted a great attention of the scientific community over the last years [Zhou, 12]. Multiple, ensemble learning models have been theoretically and empirically shown to provide significantly better performance than single weak learners, especially while dealing with high dimensional, complex regression and classification problems [Brazdil, 09], [Okun, 08]. Adaptive hybrid systems has become essential in computational intelligence and soft computing, as being able to deal with evolving components [Lughofer, 11], non-stationary environments [Sayed-Mouchaweh, 12] and concept drift (as presented in the first paper of this special issue, see below). Another main reason for their popularity is the high complementary of its components. The integration of the basic technologies into hybrid machine learning solutions [Cios, 02] facilitate more intelligent search and reasoning methods that match various domain knowledge with empirical data to solve advanced and complex problems [Sun, 00].

Both ensemble models and hybrid methods make use of the information fusion concept but in slightly different way. In case of ensemble classifiers, multiple but homogeneous, weak models are combined (e.g., see [Kajdanowicz, 10]), typically at the level of their individual output, using various merging methods, which can be grouped into fixed (e.g., majority voting), and trained combiners (e.g., decision templates) [Kuncheva, 04]. Hybrid methods, in turn, combine completely different, heterogeneous machine learning approaches [Castillo, 07], [Corchado, 10]. They both, however, may considerably improve quality of reasoning and boost adaptivity of the entire solutions. For that reason, ensemble and hybrid methods have found application in numerous real word problems ranging from person recognition, through medical diagnosis, bioinformationcs, recommender systems and text/music classification to financial forecasting [Castillo, 07], [Okun, 11], [Bergstra, 06], [Kempa, 11].

This special issue is the third one in the series of annual special issues on hybrid and ensemble methods in machine learning published by prestige scientific JCR-listed

journals after the following editions of the corresponding special sessions at the Asian Conference on Intelligent Information and Database Systems (ACIIDS). The first one appeared in *New Generation Computing*, Vol. 29, No. 3, in 2011, while the second one was published by *International Journal of Applied Mathematics and Computer Science* as a special section in Vol. 22, No. 4, 2012.

The recent special issue includes seven papers devoted to hybrid and ensemble methods as well as their application to classification and forecasting problems. It mainly originates from the *Third Special Session on Multiple Model Approach to Machine Learning* (MMAML 2012) organized by the guest editors at the *Fourth Asian Conference on Intelligent Information and Database Systems* (ACIIDS 2012), which was held in Kaohsiung, Taiwan, in March 2012. In total, ten papers were nominated by the reviewers and finally designated for oral presentation at the special session. Afterwards, the authors of some selected papers were invited to submit significantly extended versions of their contributions. Simultaneously, an open call for papers was distributed among relevant scientific community what attracted several authors. Consequently, twelve submissions were received to the current special issue. After a thorough review process, only seven of them were finally considered by the guest editors and the journal editor to become a part of the issue.

The seven accepted contributions can be classified into two different groups within the wide area of the design and application of hybrid and ensemble methods for machine learning. First four of them contain proposals of new fundamental methods, which are independent from their application area and do not require any specific domain knowledge. Their experiments studies are carried out on common reference databases widely known in machine learning to validate their correctness and compare to other known approaches. The next three papers also present new solutions but they are placed in the concrete and real application context. The datasets used for evaluation are specific and the methods proposed not directly may be applied in other domains.

Piotr Sobolewski and Michał Woźniak faced with concept drift that means the problem of significant changes in statistical properties of the target variables usually caused by some hidden and unknown features making the classification models less accurate over course of time. Detection of concept drift is very important in real dynamic environments since it may be a hint to trigger classification model reconstruction. In the contribution entitled *"Concept Drift Detection and Model Selection with Simulated Recurrence and Ensembles of Statistical Detectors"*, the authors focus on detection of virtual concept drift using unsupervised learning based on knowledge about the possible data distributions that may occur in the data stream; without any knowledge about real class labels. A priori distribution patters are treated as the known concepts, among which changes are being detected. The authors have developed their own method called simulated recurrence based on majority voting ensembles on results of statistical tests for distributions of known features. As an additional benefit, the concept detection makes the selection of the right classification model easier since a separate model may be pre-assigned to each concept.

In the second paper, entitled *"Improving Accuracy of Decision Trees Using Clustering Techniques"*, Javier Torres-Niño et al. extend fundamental classification method – decision trees by combination unsupervised and supervised machine

learning, i.e. clustering and classification. Additionally, they utilize a third component, which goal is to adjust clustering parameters. First, the predicted class attribute is removed before clustering and the number of instances in the majority class is calculated and compared with a given threshold to determine whether the instances in the entire cluster are treated as classified or not. The instances from the non-classified cluster are used to learn the decision tree. It means that clustering is performed in order to pre-classify instances based on appropriate parameters (thresholds) and popularities of classes in individual clusters. This method may be used to reduce the number of instances in the learning process what may be useful for large datasets. The authors experimentally verified various editions of their hybrid method.

Another, third fundamental research contribution entitled *"Boosting-based Multi-label Classication"* by Tomasz Kajdanowicz and Przemysław Kazienko provides a new method for the complex machine learning problem – multi-label classification, in which every instance can be independently assigned with many class labels simultaneously. The problem becomes especially demanding in case of larger output space – with many possible subsets of the class label set. The method is derived from the general boosting concept adapted to the multi-label environment. The profile of the described AdaBoostSeq algorithm has been experimentally verified on six reference datasets and three distinct base classifiers especially with respect to its robustness: for different input spaces – various numbers of input features as well as different output spaces – various numbers of distinct class labels and as a result various quantity of their power set.

Chun-Wei Lin et al. propose a new iMFFP-tree algorithm to extract fuzzy association rules in their paper entitled *"An Integrated MFFP-tree Algorithm for Mining Global Fuzzy Rules from Distributed Databases"*. Its main feature is its ability to process and integrate multiple source, local databases. It has been achieved by means of integration of many local fuzzy regions and tree branches into one coherent multiple fuzzy frequent pattern tree (MFFP-tree). It enables the authors to generate more complete global association rules, also preserving their local equivalences. The algorithm was experimentally analysed and compared against other existing approaches.

The second set of more application-oriented papers starts with the contribution entitled *"Evolutionary Fuzzy System Ensemble Approach to Model Real Estate Market based on Data Stream Exploration"* by Bogdan Trawiński. Even though the main paper focus is on predictions for the evolving real estate market, the solutions proposed may also be applied in other domains. The crucial idea behind the approach is to build a fuzzy model from the chunks of data obtained from the incoming data stream. The author utilizes evolutionary fuzzy approach coupled with the ensemble technique to explore dynamic environments – data streams. He periodically creates a new genetic fuzzy system (GFS) and merges it with the previous partially aged GFSs in order to obtain a comprehensive ensemble. The properties of the method were extensively tested on real data sets.

Thi Nhan Le et al. in their manuscript *"A Semi-Supervised Ensemble Learning Method for Finding Discriminative Motifs and Its Application"* worked out a semi-supervised learning method to discover discriminative motifs from rarely labelled biomedical sequences, and used the proposed method to distinguish difference classes

of sequences of NS5A protein regions for the Hepatitis C virus. They presented a method called E-SLUPC, which extends existing SLUPC approach by means of ensembles. It extracts motifs named *discriminative one occurrence per sequence* (DMOPS) and then applies a motif matching algorithm for label assigning. Experiments demonstrated the feasibility of the method in distinguishing real labelled dataset from the Los Amalos HCV database and unlabelled dataset from HVDB and GenBank.

In the last contribution to the special issue, Xuan Hau Pham et al. introduce a new hybrid methodology for correction processes supported by expert recommendations. In their paper *"Integrating Multiple Experts for Correction Process in Interactive Recommendation Systems"*, they demonstrate how to make the system more reliable by correction of user ratings made by the recommended experts. For that purpose, they have built the consensual recommendation framework to determine incorrect ratings, suggest experts based on user and expert ratings, relationships between users and experts as well as the consensus of experts using the convergent rating interval. The authors validated their solution on two real data sets.

Finally, the guest editors of this special issue would like to thank all the authors for their high quality contributions and twenty four independent reviewers from fourteen countries 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 the Editors-in-Chief of the Journal of Universal Computer Science, for providing us with the opportunity to edit this special issue.

Guest Editors
Przemysław Kazienko
Edwin Lughofer
Bogdan Trawiński

References

[Bergstra, 06] Bergstra, J., Casagrande, N., Erhan, D., Eck, D., Kegl, B.: Aggregate features and Ada-Boost for music classification, Machine Learning, vol. 65, pp. 473–484 (2006)

[Brazdil, 09] Brazdil, P., Giraud-Carrier, C., Soares, C.: Metalearning: Applications to Data Mining, Springer Verlag, Berlin Heidelberg (2009)

[Castillo, 07] Castillo, O., Melin, P., Pedrycz, W.: Hybrid Intelligent Systems: Analysis and Design (Studies in Fuzziness and Soft Computing), Springer, Berlin Heidelberg (2007)

[Cios, 02] Cios, K.J., Kurgan, L.A.: Hybrid Inductive Machine Learning: An Overview of CLIP Algorithms, in: Jain, L.C., Kacprzyk, J. (Eds), New Learning Paradigms in Soft Computing, pp. 276–322 (2002)

[Corchado, 10] Corchado, E., Abraham, A., de Carvalho, A.: Editorial: Hybrid intelligent algorithms and applications, Information Sciences, vol. 180 (14), pp. 2633–2634 (2010)

[Kajdanowicz, 10] Kajdanowicz, T., Kazienko, P., Kraszewski, J.: Boosting algorithm with sequence-loss cost function for structured prediction, HAIS 2010, LNAI 6076, pp. 573–580, Springer, Heidelberg (2010)

[Kempa, 11] Kempa, O., Lasota, T., Telec, Z., Trawiński, B.: Investigation of bagging ensembles of genetic neural networks and fuzzy systems for real estate appraisal. N.T. Nguyen et al. (Eds.): ACIIDS 2011, LNAI 6592, pp. 323–332, Springer, Heidelberg (2011)

[Kuncheva, 04] Kuncheva, L.: Combining pattern classifiers: Methods and algorithms, Wiley-Interscience (John Wiley \& Sons), Southern Gate, Chichester, West Sussex, England (2004)

[Lughofer, 11] Lughofer, E.: Evolving Fuzzy Systems – Methodologies, Advanced Concepts and Applications, Springer, Berlin Heidelberg (2011)

[Okun, 08] Okun, O., Valentini, G. (Eds.): Supervised and Unsupervised Ensemble Methods and their Applications (Studies in Computational Intelligence, vol. 126), Springer, Heidelberg (2008)

[Okun, 11] Okun, O.: Feature Selection and Ensemble Methods for Bioinformatics: Algorithmic Classification and Implementations, IGI Global, Hershy, PA, U.S.A. (2011)

[Sayed-Mouchaweh, 12] Sayed-Mouchaweh, M., Lughofer, E. (Eds.): Learning in Non-Stationary Environments: Methods and Applications, Springer, New York (2012)

[Sun, 00] Sun, R., Wermter, S.: Hybrid Neural Systems, Springer, Heidelberg New York (2000)

[Zhou, 12] Zhou, Z.-H.: Ensemble Methods: Foundations and Algorithms (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series), Chapman & Hall / CRC, Boca Raton, FL (2012)