ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

An Exploration of Research Directions in Machine Ensemble Theory and Applications

A. R. Figueiras-Vidal[1] and L. Rokach[2] *

1- Universidad Carlos III de Madrid,
Dept. of Signal Theory and Communications,
Avda. de la Universidad, 30, 28911, Leganés, Madrid, Spain

2- Ben-Gurion University of the Negev,
Dept. of Information Systems Engineering,
Beer-Sheva 84105, Israel.

Abstract. A concise overview of the fundamentals and the main types of machine ensembles serves to propose a structured perspective for the papers that are included in this special session. The subsequent brief discussion of the works, emphasizing their principal contributions, permits an extraction of a series of suggestions for further research in the fruitful area of ensemble learning.

1 Introduction

Machine ensembles are drawing an increasing attention, and are swiftly becoming the method of choice for supervised learning, because they offer high performance without requiring a very delicate design and a huge training effort. This is due to their conception as an aggregation of the outputs of relatively simple component machines, or units. As a consequence, an easy design of the units and a not too difficult sizing of the ensemble are allowed, even if the units and the aggregation scheme are jointly trained.

Adequately combining the modest capabilities of a number of machines allows for a high representational power without the great obstacle of selecting an appropriate (global) architecture or training it, as explained in [1]. The connections of this approach to general concepts, such as collective intelligence [2], which in the machine learning context can be interpreted as the power of weak learners [3], is clearly discussed in [4]. We can say that machine ensembles are a good combination of Occam's Razor, which prefers simple models for explanatory purposes, and Epicure's Indifference Principle, which keeps all useful models. We think that this conceptual perspective could contribute to an opening of new avenues for machine ensemble design, and also for other objectives, as we will explain later.

Of course, we do not pretend here to present a complete overview of machine ensembles, not only due to space constraints, but because there are several resourceful publications available [1][2][5][6] that already cover this topic. In the remainder of this tutorial, we first introduce a (commented) general taxonomic skeleton for machine ensembles, emphasizing some concepts and approaches we

*ARFV work is partially supported by Spain Government grant TIN2011-24533

consider relevant. We then introduce the papers which are included in this ESANN'2012 special session by locating them in the skeleton and remarking on their perspectives and main contributions. Finally, we close with suggestions for a number of potential research directions which stem from our observations of the issues addressed by these papers.

2 Types of machine ensembles

According to our knowledge, the oldest example of a machine ensemble is Selfridge's Pandemonium [7] in which a number of simple units compete for presence in the final decision according to their confidence in their own results. This scheme originated much work on committees, a basic type of ensembles in which units are first trained to solve a problem, forcing diversity among them, and subsequently their outputs are aggregated in an appropriate manner.

Diversity can be induced in many different ways such as different architectures, costs, training algorithms, input variables, training examples, etc. Among ensembles of this kind, Breiman's bagging [8], which uses bootstrap to train machine units, and its particular form Random Forests (RFs) [9] which in general designs trees with different sets of samples of reduced dimensions, are well-known due to their easy training and aggregation (usually a direct method such as the majority vote) and their good performance.

The so-called experts can be considered as the opposite approach, consisting of a series of machines which are trained for a part of the training examples (regions of the observation space), and individually selected to deal with new samples according to their expertise domain. The difficulty in selecting adequate regions for each expert and in applying appropriate training modes should not be overlooked.

Naturally, the next step is to combine, in some sense, the above approaches. This leads to what we can call collaborative ensembles. Mixtures of Experts (MoEs), first proposed in [10], are ensembles in which experts' outputs are combined in a soft manner according to weights coming from a gate. MoEs show a radical difference with respect to the above types of ensembles; the units (experts) and the gate are jointly trained (in this particular case, by means of algorithms which maximize the sample likelihood of the output, which is seen as a Gaussian mixture for regression and the exponential version of binary random variables for classification problems).

Valiant's ideas produced the first machine ensemble result [11] in a sequential filtering form, followed by the seminal contribution of boosting [12],[13] for classification purposes. The underlying idea is to sequentially train weak learners, forcing each new unit to pay more attention to the samples which produce more difficulties to be learned, and to sequentially aggregate them by means of linear combination. The surrogate cost, which was used in [12],[13], an exponential form of the margin cost (df, where d is the desired result and f the output), allows optimization of both units and a combination of coefficients in an easy way, though, this is not strictly necessary in order to obtain good designs [14]. There are also boosting modifications for regression purposes (which do not offer impressive results), as well as other sequential algorithms, such as Negative Correlation Learning (NCL) [15], which forces diversity by means of penalizing the correlation among the outputs.

Boosting has demonstrated a remarkable (and perhaps unique) resistance to overfitting, although they are affected by the presence of outliers or high levels of noise. Moreover, many modified boosting algorithms reduce this sensitivity.

The success of boosting ensembles seems to have reduced the attention that MoEs and other joint (of units and aggregation) designs receive. However, in our opinion, trying to combine different approaches to design machine ensembles is a good route to introduce diversity and increase representation capabilities. For example, using gates to aggregate boosting learners can help alleviate overfitting problems and, simultaneously, attain more expressive power, as the results of a first design of this form [16] indicate. Even a simple reorganization of MoEs [17] leads to compact architectures that offer excellent classification performance when trained by means of Support Vector (SV) [18] algorithms. We assert that combining global and local capabilities is an essential element in order to obtain competitive designs.

In this short section we presented a coarse framework for ensemble learning. The next section introduces the contributions to the special session.

3 Introducing the contributions to the special session

Four of the contributions deal with different aspects of what we will call committees.

Paper [19] analyzes a committee whose units are trained with randomized versions of a learning algorithm. Consequently, their outputs can be considered statistically independent for different examples. Thus, a statistical test can be designed to identify instances that are close to decision borders. Paying attention to these instances is very important to improve the performance of any kind of classification machine, as the extensive experience in training example selection and weighting indicates since the pioneering work of Hart [20]. It is also relevant to select kernels for SV Machines and related architectures, as discussed in [21] and subsequent contributions.

Work [22] proposes to construct committees of Extreme Learning Machines (ELMs), that are an easy and efficient method for designing traditional forms of Neural Networks (NNs), such as Single Hidden Layer Perceptrons (SHLPs), by means of regularized linear combinations. Standard ELM designs have a limited performance due to the random design of their units, however, the authors correctly state that SHLPs show enough complementary diversity and conclude that an appropriate aggregation of their outputs will provide powerful committees which do not require a high design effort. Experimental results support their point of view. Let us add that, in fact, the ELM design itself can be considered a committee construction problem, and that many alternatives to their standard versions do exist.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Paper [23] defines a method for selecting the components of an ensemble to solve multi-class classification problems by applying a well principled measure that includes diversity. The experimental results are again a proof of the fundamental importance of diversity in machine ensembles.

From [19],[22],[23], one can conclude that there are some sources of diversity that have not been sufficiently explored and that combining them, using different diversity measures, and even introducing diversity in the aggregation processes, can lead to improved committee designs.

The fourth paper of this group [24] shows that (apparent) oversizing of RFs is important in order to get stability in both classification and feature selection. We should note that RFs, although consisting of trees, are still considered to be uncomprehensible, but are useful for variable selection. It could be interesting to check whether other ensemble architectures present this advantage.

Poster [25] proposes a reduction of the target matrix of a multi-class classification problem, and a design of kernels by means of a boosting-type algorithm applied to a linear combination of inner products of the outputs of some base learners. This second process is a good example of how boosting ideas can be applied to achieve very diverse objectives. As we mentioned before, boosting has extended to an extensive variety of algorithms that apply very diverse progressive sample weighting and aggregation modes. In our opinion, it is important to get practical guidance on which modes are adequate to face which specific problems. On the other hand, this diversity of boosting algorithms could be used to construct ensembles of boosting ensembles.

Paper [26] comes from an area of application which is attracting more and more interest, namely, distributed learning. The significant deployment of communication facilities produces a pervasive expansion of many types of distributed systems, an important fact when considering that distribution is a primary source of diversity. The authors of [26] introduce a "collective" algorithm based on error gradient diffusion which preserves data privacy and offers a very definitive advantage, with respect to previous approaches, of maintaining the ability to track environmental changes while at the same time providing a good degree of convergence for stationary situations. Algorithm 1 in [24] constitutes a key contribution to the introduction of the obviously necessary (real-time) adaptation capabilities in applications which are inherently non-stationary. However, at the same time, the notion of interchanging learning orientations, such as error gradients, and their use by the learners, appears to be a new possibility for construction of general collaborative ensembles.

4 Some suggestions for further research

The six works included in this special session would merit more extensive and deeper discussion, however, space and personal knowledge limitations allow us just to take a few general recommendations from their analysis of some fundamental aspects, with the hope of being useful for future research approaches to the wonderful world of machine ensembles.

1. Learning machines can provide useful information and guidance for designing other machines or for their learning process. An interesting research subject might be on which kind of appropriate information is available and what kind of applications are associated with. Note that boosting is an approach of this type, but there are interesting alternatives, such as I-votes [27], which progressively construct simple learners by means of an appropriate selection of samples, or diffusion algorithms. Of course, combining these possibilities is also interesting.

2. Complementary diversity is a key element to build ensembles, both committees and collaborative schemes. There are diversity sources that have not been fully explored such as randomness, and even task diversity [28]. Places and criteria to use diversity also need a deeper exploration. For example, aggregation diversity is seldom considered. Again, combinations can provide advantages.

   Note, that 1 and 2 call for a creative revision of existing ensemble design techniques.

3. There are some general problems that have not been fully addressed in the machine ensemble literature. Among them, we feel that multi-class, cost-sensitive, sparsity-aware and non-stationary learning are particularly important.

4. Distributed learning is an area of research which has an increasing importance and requires machine ensembles. An adequate revision of information interchange options, plus the introduction of incremental or even real-time learning, are critical questions to be addressed in order to find practical solutions to many problems. Note that, additionally, this kind of knowledge could be important to carry out an analysis of different aspects of the human collective intelligence.

5. Following our human collective intelligence reference, we will close these lines with an invitation to consider a challenging question: Since human and machine decision making are essentially diverse, how can we combine them to get better decisions and to favor human knowledge acquisition?

References

[1] L. I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms.* Wiley, Hoboken, NJ, 2004.

[2] J. Surowiecki. *The Wisdom of Crowds.* Doubleday, New York, NY, 2004.

[3] L. G. Valiant. A theory of the learnable. *Communications ACM*, 27:1134–1142, 1984.

[4] L. Rokach. *Pattern classification using ensemble methods.* World Scientific, Singapore, 2010.

[5] A.J.C. Sharkey, editor. *Combining artificial neural nets: ensemble and modular multinet systems.* Springer, London,UK, 1999.

[6] G. Seni and J. Elder. *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions.* Morgan and Claypool, San Rafael, CA, 2010.

[7] O. G. Selfridge and U. Neisser. Pattern recognition by machine. *Scientific American*, 203(2):60–68, 1960.

[8] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[9] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[10] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.

[11] R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990.

[12] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[13] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37:297–336, 1999.

[14] L. Breiman. Arcing classifiers. *The Annals of Statistics*, 26(3):801–824, 1998.

[15] Y. Liu and X. Yao. Ensemble learning via negative correlation. *Neural Networks*, 12:1399–1404, 1999.

[16] E. Mayhua-López, V. Gómez-Verdejo, and A. R. Figueiras-Vidal. Real adaboost with gate controlled fusion. Submmited to *IEEE Trans. Neural Networks and Learning Systems*, 2012.

[17] A. Omari and A. R. Figueiras-Vidal. Feature combiners with gate generated weights for classification. Submmited to *IEEE Trans. Neural Networks*, 2011.

[18] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pp. 144–152, Pittsburgh, PA, 1992.

[19] D. Hernández-Lobato, G. Martínez-Muñoz, and A. Suárez. On the independence of the individual predictions in parallel randomized ensembles. In *ESANN'2012 Special Session on Machine Ensembles*, 2012.

[20] P. E. Hart. The condensed nearest neighbor rule. *IEEE Trans. on Information Theory*, 14:515–516, 1968.

[21] A. Lyhyaoui, M. Martínez, I. Mora, M. Vázquez, J. L. Sancho, and A. R. Figueiras-Vidal. Sample selection via clustering to construct support vector-like classifiers. *IEEE Trans. on Neural Networks*, 10(6):1474–1481, 1999.

[22] P. Escandell-Montero et al. Regularized committee of extreme learning machines for regression problems. In *ESANN'2012 Special Session on Machine Ensembles*, 2012.

[23] L. Chekina, L. Rokach, and B. Shapira. Introducing diversity among the models of multi-label classification ensemble. In *ESANN'2012 Special Session on Machine Ensembles*, 2012.

[24] J. Paul, M. Verleysen, and P. Dupont. The stability of feature selection and class prediction from ensemble tree clasifiers. In *ESANN'2012 Special Session on Machine Ensembles*, 2012.

[25] A. Lechervy, P. H. Gosselin, and F. Precioso. Linear kernel combination using boosting. In *ESANN'2012 Special Session on Machine Ensembles*, 2012.

[26] Z. Towfic, J. Chen, and A. H. Sayed. Distributed learning via difussion adaptation with application to ensemble learning. In *ESANN'2012 Special Session on Machine Ensembles*, 2012.

[27] L. Breiman. Pasting small votes for classification in large databases and on-line. *Machine Learning*, 36:85–103, 1999.

[28] R. Caruana. *Multitask Learning*. Ph. D. Diss., School of Computer Science, Carnegie-Mellon University, 1997.