Fellowship Scientific Report

Fellow: Christian Gagné
Visited Location : University of Lausanne (Switzerland)
Duration of Visit: 9 months

I - Scientific activity

The researches done at the University of Lausanne for the second part of the fellowship were in continuity with those started in France. Indeed, it was agreed with Marc Schoenauer of the INRIA (France) and Marco Tomassini of the University of Lausanne (Switzerland), supervisors of the fellow, on a common general research program spanning the whole duration of the fellowship.

The research projects are at the interface between machine learning and evolutionary algorithms. On one hand, machine learning consists in developing intelligent systems showing learning capabilities. This learning is usually done by an automatic analysis of data that are observations of the phenomenon to model. On the other hand, evolutionary algorithms are a family of black-box population-based global optimization algorithms inspired by natural evolution. Evolutionary algorithms have some properties that can be of great utility for developing machine learning systems. Symmetrically, the strong statistical roots of machine learning can help improving the current evolutionary algorithms on different aspects such parameters tuning and testing methodologies.

Researches done during the first months in Switzerland focused mainly on improving a system for the evolution of kernel functions by genetic programming, an evolutionary algorithm used for automatic program induction. Kernel functions are similarity measures in some machine learning algorithms such the Support Vector Machines (SVM). Although there are some widely known kernel functions (e.g. Gaussian kernel, polynomial kernel), there is no known method for selecting or designing kernel functions adapted to the problem at hand. The choice of a kernel appropriated to a given task usually depends on the good understanding of the problem domain by human experts and some experimentations for tuning the kernel's hyper-parameters.

The system developed during the fellowship aims at automate this selection of kernel by using genetic programming to generate kernel functions appropriated for a given data set. A first version of the system has been presented in [2], with interesting results on benchmark data sets for classification. Further works have been done to improve performance of the system, and to generalize the approach to contexts other than classification. This work can be of great and general incidence on machine learning. Indeed, kernel-based learning appears now to be a major breakthrough of the field of machine learning in the last 15 years, and is now applied in several common learning contexts (e.g. classification, clustering, dimensionality reduction). Proposing a general method for kernel selection is an answer to some of the problems raised by these methods. In the months following the end of the fellowship, it is planned to prepare a paper presenting the improved version of the system for the genetic programming of kernel functions, which will be submitted to a major journal on machine learning or evolutionary computation.

Collaboration has been enacted with the Institute of Geomatics and Risk Analysis of the University of Lausanne in order to apply the system for the evolution of kernel functions to geospatial data. A technological transfer has already been done and the collaborating peoples in geomatics, who are currently doing several experimentations to test the system on their data. This collaboration may lead to one or some scientific publications in geomatics and machine learning.

For the last months of the fellowships in Switzerland, investigations have been conducted on the inclusion of ensemble learning in evolutionary computation. Ensemble learning is based on the idea that using committee of classifiers can lead to more robust, thus more general classification systems, compared to single classifiers. Robustness and generalization are major objectives of supervised machine learning. Approaches based on committee are highly relevant for learning with evolutionary algorithms. Indeed, contrary to usual machine learning algorithms which generate a single solution as result of their application, evolutionary algorithms produce as final result a population of solutions. Thus, instead of using only the best solution of the evolution as final result, it can be more appropriate to use an ensemble of solutions that would perform better that the best-of-run solution. This would allow an efficient use of the numerous solutions provided by a single evolution.

A general methodology for evolutionary ensemble learning has been developed during the fellowship. Its includes a special fitness function that aims at optimizing the diversity of the output given by the evolved solutions and a procedure to build an ensemble of evolved classifiers from the final population. Experimental results show significant improvement of classification rate over using single best-of-run evolved classifier. A conference paper will be prepared in the weeks following the end of the fellowship for a submission to an international conference on evolutionary algorithms.

Collaboration with peoples from the INRIA to diverse research projects on evolutionary computation have been done during the fellowship. The most notable collaboration is on a paper for the analysis of the bloat phenomenon, common in genetic programming, using elements from statistical machine learning. The collaboration in the context of this fellowship was essentially to provide several experimental results to back the theory previously developed. This research is the object of a paper [1] that have been submitted to the international journal *Genetic Programming and Evolvable Machines*. Another collaboration, with the same peoples of the INRIA, is on the use of quasi-random numbers (common in Monte-Carlo simulations) to replace the pseudo-random number generators of classical evolutionary algorithms. This project may also lead to production of scientific papers in a near future. A project has also been initiated in the context of this fellowship for modifying classical genetic programming by including some idea taken from self-adapting evolution strategies (another evolutionary algorithm flavor). These modifications imply using a variable-size small population and mutations simulating some kind of stochastic hill-climbing mutations. Experimentations on this project are still on-going.

Some attention has been given during the fellowship to maintaining, developing, and promoting the software tool “Open BEAGLE”. This software, developed during the master and PhD of the fellow, is an open source, generic C++ library for evolutionary computation, freely available on the Web. It is now widely known and used in the evolutionary computation community. A short article on Open BEAGLE has been recently published in the ACM Special Interest Group on Genetic and Evolutionary Computation (SIGEVO) newsletter [4]. A new version of the software has also been released at the end of the fellowship.

Finally, the fellow attended to two major scientific conferences on evolutionary computation during the visit to Switzerland, namely EuroGP 2006 in Budapest and PPSN IX in Reykjavik. At both of these conferences, the fellow did an oral presentation of accepted papers [2,5]. The fellow also did a presentation of the system for the evolution of kernels to a machine learning group of the IDIAP, in Martigny (Switzerland), as well as presentation to two internal seminars at the University of Lausanne. The fellow also reviewed several papers as member of the program committee for the GECCO 2006, IEEE-CEC 2006, IJCAI 2007, and EuroGP 2007 conferences.

In conclusion, this fellowship leads to many fructuous collaborations with researchers from the INRIA and the University of Lausanne. Different projects, all closely related to evolutionary computation and machine learning, have resulted in the production of several scientific papers during the fellowship or in the months that will follow.

II- Publications during the fellowship

[1] Nur Merve Amil, Nicolas Bredeche, Christian Gagné, Sylvain Gelly, Marc Schoenauer, and Olivier Teytaud, "A Statistical Learning Perspective of Genetic Programming", submitted to *Genetic Programming and Evolvable Machines*, October 2006.

Abstract: *Code bloat, the excessive increase of code size, is an important issue in Genetic Programming (GP). This paper proposes a theoretical analysis of code bloat in GP from the perspective of statistical learning theory, a well grounded mathematical toolbox for machine learning. By computing the Vapnik-Chervonenkis dimension of the family of programs that can be inferred by a specific setting of GP, it is proved that a parsimonious fitness ensures universal consistency. This mean that the empirical error minimization allows converge to the best possible error when the number of test cases goes to infinity. However, it is also proved that the standard method consisting in putting a hard limit on the program size still results in programs of infinitely increasing size in function of their accuracy. It is also shown that cross-validation or hold-out for choosing the complexity level that optimizes the error rate in generalization also leads to bloat. So a more complicated modification of the fitness is proposed in order to avoid unnecessary bloat while nevertheless preserving universal consistency.*

[2] Christian Gagné, Marc Schoenauer, Michèle Sebag, and Marco Tomassini, "Genetic Programming for Kernel-based Learning with Co-evolving Subsets Selection", *In Proc. of Parallel Problem Solving from Nature (PPSN IX)*, Reykjavik (Iceland), September 9-13 2006, p. 1008-1017.

Abstract: *Support Vector Machines (SVMs) are well-established Machine Learning (ML) algorithms. They rely on the fact that i) linear learning can be formalized as a well-posed optimization problem; ii) nonlinear learning can be brought into linear learning thanks to the kernel trick and the mapping of the initial search space onto a high dimensional feature space. The kernel is designed by the ML expert and it governs the efficiency of the SVM approach. In this paper, a new approach for the automatic design of kernels by Genetic Programming, called the Evolutionary Kernel Machine (EKM), is presented. EKM combines a well-founded fitness function inspired from the margin criterion, and a co-evolution framework ensuring the computational scalability of the approach. Empirical validation on standard ML benchmark demonstrates that EKM is competitive using state-of-the-art SVMs with tuned hyper-parameters.*

[3] Sylvain Gelly, Olivier Teytaud, and Christian Gagné, "Resource-Aware Parameterizations of EDA", *In Proc. of the IEEE Congress on Evolutionary Computations (IEEE-CEC 2006)*, July 16-21, 2006, Vancouver, BC (Canada).

Abstract: *This paper presents a framework for the theoretical analysis of Estimation of Distribution Algorithms (EDA). Using this framework, derived from the VC-theory, we proposed non-asymptotic bounds which depend on: 1) the population size, 2) the selection*

rate, 3) the families of distributions used for the modeling, 4) the dimension, and 5) the number of iterations. To validate these results, optimization algorithms are applied to a context where bounds on resources are crucial, namely Design of Experiments, that is a black-box optimization with very few function value evaluations.

[4] Christian Gagné and Marc Parizeau, "Open BEAGLE, A C++ Framework for your Favorite Evolutionary Algorithm", SIGEVOlution, Vol. 1, no. 1, p. 12-14, April 2006.

[5] Christian Gagné, Marc Schoenauer, Marc Parizeau, and Marco Tomassini, "Genetic Programming, Validation Sets, and Parsimony Pressure", *In Proc. of the European Conference on Genetic Programming (EuroGP 2006)*, vol. 3905 of LNCS, pp. 109-120, Springer, April 10-12 2006, Budapest (Hungary).

Abstract: *Fitness functions based on test cases are very common in Genetic Programming (GP). This process can be assimilated to a learning task, with the inference of models from a limited number of samples. This paper is an investigation on two methods to improve generalization in GP-based learning: 1) the selection of the best-of-run individuals using a three data sets methodology, and 2) the application of parsimony pressure in order to reduce the complexity of the solutions. Results using GP in a binary classification setup show that while the accuracy on the test sets is preserved, with less variances compared to baseline results, the mean tree size obtained with the tested methods is significantly reduced.*

III -Attended Seminars, Workshops, and Conferences

- PPSN IX, *the 9th International Conference on Parallel Problem Solving from Nature*, September 9-13, 2006, Reykjavik (Iceland).

- EuroGP 2006, *the 9th European Conference on Genetic Programming*, April 10-12, 2006, Budapest (Hungary).