Fellowship Scientific Report

Fellow: Christian Gagné
Visited Location : INRIA (France)
Duration of Visit: 9 months

I - Scientific activity

The researches conducted during the postdoctoral fellowship 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.

The French part of the fellowship started by a study on the use of machine learning methodologies for genetic programming (an evolutionary algorithm). More specifically, it aims at illustrating the use of a validation data set for selecting best-of-runs individuals of genetic programming evolutions that generalize well. Results obtained with the proposed methodology show a significant size reduction of the best-of-runs individuals compared to the naive approach, while maintaining the performance accuracy. A paper presenting this study has been accepted for presentation at the EuroGP 2006 conference.

The Support Vector Machines (SVM) classifiers represent certainly one of the most important breakthrough in machine learning of the last decade. SVM consist in kernel-based learning machine based on strong statistical principles. Kernel-based learning rely on the idea of transforming an ill-posed learning problem in the input space into an well-posed, hopefully linearly separable, learning problem in an high dimension space. This mapping into an high dimension space is done using with an arbitrary function, the so-called kernel function. But there is no deterministic method to select and parameterized the kernel function, so the choices made for this kernel functions can greatly influence the results obtained. In the present postdoctoral fellowship, we explored the idea of using genetic programming to design a data-specific kernel function. This researches lead to an evolutionary equivalent to the SVM, the Evolutionary Kernel Machine, which co-evolve the kernel function optimization with some data subset selection for nearest neighbor classification. Results obtained are very encouraging, showing that the proposed system is competitive with classical approaches. A paper on the system has been submitted to the PPSN conference. Most of the second part of the postdoctoral fellowship in Switzerland will deal with improving and analyzing this Evolutionary Kernel Machine.

The stay in France for the postdoctoral fellowship has also lead to several fruitful collaborations. Indeed, collaboration on the use of evolutionary algorithms for optimization problems with very limited number of cost function evaluation has been accepted for a publication at the CEC 2006 conference. Several other on-going works should lead to papers submission in international conferences and journals. An example of is a paper on the bloat phenomenon in genetic programming using elements from the computational learning theory that will be submitted to an international journal in the next months. Another example will be to study the effect of using quasi-random numbers, common in Monte-Carlo simulations, for real-valued optimization with

evolutionary algorithms. There is also a strong collaborations with my PhD adviser at the Université Laval (Québec, Canada), in order to complete the writing of several papers that stems from my thesis. Finally, some time is given to support the developments of the open source Open BEAGLE library, which is a programming environment for evolutionary algorithms researches we started developing several years ago.

II- Publications during the fellowship

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.*

Sylvain Gelly, Olivier Teytaud, and Christian Gagné, "Resource-Aware Parameterizations of EDA", *Accepted for publication at the IEEE Congress on Evolutionary Computations (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.*

Christian Gagné, Marc Schoenauer, Michèle Sebag, and Marco Tomassini, *"*Genetic Programming for Kernel-based Learning with Co-evolving Subsets Selection"*, Submitted to the International Conference on Parallel Problem Solving from Nature (PPSN IX)*, September 9-13 2006, Reykjavik (Iceland).

Abstract: *Support Vector Machines (SVM) are well-established machine learning systems. They are characterized by: 1) a transformation by a kernel function of the feature space into an high dimension space, 2) the use of linear classifier working in this high dimension space, parameterized by quadratic programming, and 3) the maximization of the margins between the separating hyper-planes and the training data. Inspired by this classifier, we propose a system named the Evolutionary Kernel Machine (EKM) that consists in: 1) the genetic programming of kernel functions adapted to the data at hand, 2) a data classification according to the nearest neighbor in the high dimension space, and 3) the maximization of a neighborhood ranking differences as fitness function. In order to reduce the computational requirements of the fitness function, the system also incorporates a co-evolutionary subset selection of the prototypes and the test cases. Results obtained show that the EKM is quite competitive compared with k-nearest neighbors and SVM classifiers.*

III -Attended Seminars, Workshops, and Conferences

EA 2005, *the 7th International Conference on Artificial Evolution*, October 26-28, 2005 - University of Lille (France).