INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team tao

Thème apprentissage et optimisation

Futurs

THEME COG

Table of contents

1. Team **1**
2. Overall Objectives **2**
2.1. Overall Objectives 2
3. Scientific Foundations **2**
3.1. Scientific Foundations 2
3.1.1. Machine learning, Data Mining, Inductive Logic Programming 2
3.1.2. Evolutionary Computation, Stochastic Optimization 3
3.1.3. Modelling and Control of Complex Systems 3
4. Application Domains **4**
4.1. Application Domains 4
5. Software **4**
5.1. Evolving Objects 4
5.2. GUIDE: A graphical interface for EO 4
5.3. World in a bottle 4
5.4. Simbad Autonomous and Evolutionary Robotics Simulator 5
5.5. PuppetMaster - Generic 3D Robotics Simulator 5
5.6. Django 6
5.7. OpenDP 6
6. New Results **7**
6.1. Fundamentals of Machine Learning, Knowledge Extraction and Data Mining 7
6.1.1. New learning criteria 7
6.1.2. Selection of Features and Patterns 8
6.1.3. Resampling algorithms 8
6.1.4. Relational learning and phase transitions 8
6.1.5. Exploration vs Exploitation and Multi-armed Bandits 9
6.2. Fundamentals of Evolutionary Computation 10
6.2.1. Convergence Analysis for Evolution Strategies 10
6.2.2. Genetic Programming 11
6.2.3. Surrogate models in Evolution Strategies 11
6.2.4. Estimation of Distribution Algorithms 12
6.3. Modelling and Control of Complex Systems 12
6.3.1. Robotics 12
6.3.2. e-Science and Grid Modelling 13
6.3.3. Evolutionary design 14
6.3.4. Inverse problems 15
6.3.5. Optimization and Identification of Complex Networks 16
6.4. Other applications 17
6.4.1. Text Mining 17
6.4.2. Large-Eddy Simulations 17
6.4.3. Train scheduling 17
6.4.4. Time-dependent planning with bounded resources 18
6.4.5. Multi-disciplinary multi-objective optimization 18
7. Contracts and Grants with Industry **18**
7.1. Contracts and Grants with Industry 18
8. Other Grants and Activities **20**
8.1. International actions 20
8.1.1. Management positions in scientific organizations 20
8.1.2. Collaborations with joint publications 20
8.2. European actions 20

8.2.1. Management positions in scientific organizations 20
8.2.2. Working groups 20
8.2.3. Collaborations with joint publications 20
8.3. National actions 20
8.3.1. Organization of conferences and scientific events 20
8.3.2. Management positions in scientific organizations 20
8.3.3. Associations 21
8.3.4. Collaborations with joint publications 21
8.4. Honors 21
8.4.1. Keynote addresses 21
9. Dissemination **21**
9.1. Animation of the scientific community 21
9.1.1. Editorial boards 21
9.1.2. Chair in Organizing Committees 22
9.1.3. Program Committee Member (international events) 22
9.1.4. Program Committee Member (national events) 22
9.1.5. Evaluation committees and invited expertise 22
9.1.6. Other evaluation activities 22
9.1.7. Summer schools, tutorials, invited seminars 22
9.2. Enseignement 23
9.2.1. Defended doctorates 23
9.2.2. Graduate courses 23
9.2.3. Other research-related teaching activities 23
10. Bibliography **23**

1. Team

TAO (Thème Apprentissage et Optimisation) is a joint project inside PCRI, including researchers from INRIA and the LRI team I & A – Inférence et Apprentissage (CNRS and University of Paris Sud), located in Orsay.

Co-Heads of project-team

Marc Schoenauer [DR INRIA, HdR]
Michèle Sebag [DR CNRS, HdR]

Permanent staff

Anne Auger [CR INRIA – since October 2006]
Nicolas Bredèche [MDC Université Paris Sud]
Philippe Caillou [MDC Université Paris Sud]
Cécile Germain [Professeur Université Paris Sud, HdR]
Olivier Teytaud [CR INRIA]

Administrative assistant

Marie-Carol Lopes [Tao/In Situ]

Ph.D. Students

Nicolas BASKIOTIS [started September 2003]
Alexandre DEVERT [started September 2005]
Lou FEDON [started September 2006]
Mary FELKIN [started September 2003]
Romaric GAUDEL [started September 2006]
Sylvain GELLY [started September 2004]
Cédric HARTLAND [started September 2005]
Thomas HEITZ [started September 2004]
Celso Yoshikazu ISHIDA [started December 2003, co-tutelle Universidade Federal do Parana - Brésil)]
Mohamed JEBALIA [started September 2004, co-tutelle Université Tunis - Tunisie)]
Fei JIANG [started November 2006 TAO-ALCHEMY]
Claire LE BARON [CIFRE Renault, started September 2005]
Julien PEREZ [started October 2006]
Vijay PRATAP SINGH [thèse IFP, defended December 2006]
Arpad RIMMEL [started September 2006]
Raymond ROS [started September 2005]
Xiangliang ZHANG [started September 2006]

Post-doctoral fellows

Cyril FURTLEHNER [since October 2006]
Ettore CAVALLARO [since November 2006]
Christian GAGNÉ [until March 2006]
Jérémie MARY [until September 2006]

Research Engineer

Bertrand CHARDON
Miguel NICOLAU [since October 2006]
Yann SEMET [until May 2006]
Damien TESSIER

Collaborators

Balasz KEGL [CR CNRS, LAL]
Yves KODRATOFF [DR CNRS, HdR]

2. Overall Objectives

2.1. Overall Objectives

Data Mining (DM) has been identified as one of the ten main challenges of the 21st century (MIT Technological Review, fev. 2001). The goal is to exploit the massive amounts of data produced in scientific labs, industrial plants, banks, hospitals or supermarkets, in order to extract valid, new and useful regularities. In other words, DM resumes the Machine Learning (ML) goal, finding (partial) models for the complex system underlying the data.

DM and ML problems can be set as optimization problems, thus leading to two possible approaches. Note that this alternative has been characterized by H. Simon (1982) as follows. *In complex real-world situations, optimization becomes approximate optimization since the description of the real-world is radically simplified until reduced to a degree of complication that the decision maker can handle. Satisficing seeks simplification in a somewhat different direction, retaining more of the detail of the real-world situation, but settling for a satisfactory, rather than approximate-best, decision.*

The first approach is to simplify the learning problem to make it tractable by standard statistical or optimization methods. The alternative approach is to preserve as much as possible the genuine complexity of the goals (yielding “interesting” models, accounting for prior knowledge): more flexible optimization approaches are therefore required, such as those offered by Evolutionary Computation.

Symmetrically, optimization techniques are increasingly used in all scientific and technological fields, from optimum design to risk assessment. Evolutionary Computation (EC) techniques, mimicking the Darwinian paradigm of natural evolution, are stochastic population-based dynamical systems that are now widely known for their robustness and flexibility, handling complex search spaces (e.g. mixed, structured, constrained representations) and non-standard optimization goals (e.g. multi-modal, multi-objective, context-sensitive), beyond the reach of standard optimization methods.

The price to pay for such properties of robustness and flexibility is twofold. On one hand, EC is tuned, mostly by trials and errors, using quite a few parameters. On the other hand, EC generates massive amounts of intermediate solutions. It is suggested that the principled exploitation of preliminary runs and intermediate solutions, through Machine Learning and Data Mining techniques, can offer sound ways of adjusting the parameters and finding short cuts in the trajectories in the search space of the dynamical system.

3. Scientific Foundations

3.1. Scientific Foundations

Abstract:

One of the goals of Machine Learning and Data Mining is to extract optimal hypotheses from (massive amounts of) data. What "optimal" means varies with the problem. The goal might be to induce useful knowledge, allowing new cases to be classified with optimal confidence (predictive data mining), or to synthesize the data into a set of understandable statements (descriptive data mining).

On the other hand, Evolutionary Computation and stochastic optimization are adapted to ill-posed optimization problems, such as involved in machine learning, data mining, identification, optimal policies, and inverse problems. However, optimization algorithms must adapt themselves to the search landscape; in other words, they need learning capabilities.

3.1.1. Machine learning, Data Mining, Inductive Logic Programming

Learning and mining are concerned with i) choosing the form of knowledge to be extracted (e.g., rules, Horn clauses, distributions, patterns, equations), referred to as hypothesis space or language ; ii) exploring this (huge) search space to find the best hypotheses it contains.

Formally, learning and mining can be cast as optimization problems under incomplete information. For instance, statistical learning can be viewed as an incomplete information game; while the player only knows some cards of the game (the training examples), the goal is to find some hypothesis with minimal expected loss (over all possible examples in the application domain). Likewise, a data mining algorithm is expected to provide the expert user with "interesting" regularities, even though in general the expert's interestingness criteria can only be discovered along the process.

New learning criteria have been investigated, related either to the structure of the hypothesis space (e.g. Bayesian nets), or to the expert's priors (e.g. ROC-based criteria, applied to medicine and bio-informatics) or preferences (e.g., multi-objective criteria for spatio-temporal data mining, applied to brain imagery).

Meanwhile, learning and mining can also be formalized as constraint satisfaction problems (CSP), particularly so for Machine Learning in First Order Logic, referred to as Inductive Logic Programming. Thorough and fruitful efforts have been made to transport the statistical computational analysis developed for CSPs, to the learning and mining disciplines. This cross-disciplinary research led to discover the presence of a phase transition in the search landscape, with far fetched consequences on the competence and scalability of the existing algorithms.

3.1.2. Evolutionary Computation, Stochastic Optimization

Considering the lack of a universal optimization algorithm, the power of an optimization algorithm is measured by its ability in acquiring and exploiting problem-specific information. Long based on heuristics, the use of prior knowledge most often results in customized representations and search spaces, specific evolution operators, and/or additional constraints. One of our long-term objectives is to develop self-adaptive operators, able to automatically detect and exploit regularities in the search space. Another objective is to investigate a principled use of prior knowledge in every level of evolutionary algorithms, ranging from the representation and the variation operators, to the selection operator and the tuning of the fitness function, and the choice of the hyper-parameters.

3.1.3. Modelling and Control of Complex Systems

In previous years, the field of Autonomous Robotics most naturally motivated the tight coupling of Learning and Optimization approaches. *A posteriori*, it appears that many key aspects of this field (size of the state and decision spaces; continuous vs discrete modelization; possibly different training and test distributions; stability of the control; etc) are relevant to the modelling and control of complex systems at large. Such links between modelling and control of autonomous complex systems have been explored along several directions:

Reinforcement learning and control
The open platform OpenDP (Section 5.7) hybridizes the standard Bellman decomposition with (i) machine learning algorithms; (ii) derivative-free and evolutionary optimization.

- As opposed to discretization or linear interpolations, ML algorithms scale up with the dimensionality and do not require the convexity hypothesis underlying e.g. dual dynamic programming;
- Evolutionary optimization is well suited for the optimization of non convex value functions.

Modelling and autonomy
The AAA study (Robea Contract: Agir, Anticiper, s'Adapter. 2002-2005) aimed at providing the robotic system with a model of itself: self-awareness is viewed as a step toward autonomous behavior. While the targeted complex system initially was a robot, the approach is now being extended to Grid Modelling (see Section 6.3.2).

Estimating and action selection
The Multi-Armed Bandit framework, concerned with the pervasive "exploration vs exploitation" dilemma, has been considered in the context of dynamic environments and large numbers of options (see Section 6.1.5).

4. Application Domains

4.1. Application Domains

Applications are described all along the text, and referenced in the contract section.

The main application domains are Robotics (see section 6.3.1), Medical Data Mining (see sections 6.1.1,6.1.2) and Inverse problems for Numerical Engineering (section 6.3.4.

5. Software

5.1. Evolving Objects

Keywords: *Evolutionary Computation*, *Object-oriented*, *Standard Template Library*.

Participant: Marc Schoenauer [correspondent].

Abstract: EO is a templates-based, ANSI-C++ compliant evolutionary computation library. It contains classes for almost any kind of evolutionary computation you might come up to - at least for the ones we could think of. It is component-based, so that if you don't find the class you need in it, it is very easy to subclass existing abstract or concrete class. EO works with main compilers, including GNU *g++* (versions 2.95 and above) and Microsoft *Visual C++* (versions 6.00 and above).

In 2006, the 1.0 version of EO has been (at last) launched, including full support for the most recent compilers under Linux and Windows.

See main page at http://eodev.sourceforge.net/

5.2. GUIDE: A graphical interface for EO

Keywords: *Evolutionary Computation*, *GUI*, *Java*, *Object-oriented*.

Participants: Marc Schoenauer [correspondent], Damien Tessier.

Abstract: GUIDE is a graphical user interface for the Open Source library EO (see above). It allows the user to describe its genome (the structure that will evolve) graphically, represented as a tree, using containers and elementary types (booleans, integers, real numbers and permutations). All representation-dependent operators (initialization, crossover and mutation) can then be defined either using default values, built bottom-up from the elementary types, or user-defined operators. Developing a prototype for a new search space involving complex structures has now become a matter of minutes.

GUIDE was programmed in JAVA by James Manley during the 6 months of his DESS stage in 2004. It is a follow-up of a previous tool developed in collaboration with Pierre Collet in the DREAM project (http://www.dcs.napier.ac.uk/~benp/dream/dream.htm).

After being linked with WEKA in 2005 (addition of a new type of genotype, compliant with Weka `.aff` files), GUIDE has not evolved in 2006. However, it has been chosen as the evolutionary basis for the EvoTest project: testing a given program means feeding it with data of a specific structure. Because the goal of EvoTest is to automatically evolve test data, we need an automatic code generator that only requires a description of the structure of the data to evolve – and this is precisely what GUIDE is doing. The main change will be to go from interactive graphical user interface to a fully automated program interface.

5.3. World in a bottle

Keywords: *OpenGL*, *Robot simulator*.

Participant: Jeremie Mary [correspondent].

Abstract: "World in a Bottle" is a robot's simulator written in C++ that takes advantage of the OpenGL library. Khepera robots are currently privileged but it is possible to easily implement other types of robots into the simulator.

- Real time 3D display of the simulation, including sensors view (display can be disable to achieve better simulation speed).
- The environment can be easily designed (walls, cylinders...), as well and saved or loaded to/from a file.
- It is possible to run as many robots as need in the simulator.
- It is possible to control the robot during the simulation through the keyboard controls.
- Simulation of IR proximity sensors both in active and passive modes.
- Simulation of 1D and 2D Cameras.
- Simulation of moving obstacles.
- It is possible to easily write a controller in C++ (C++ tutorials are available).
- It is also possible to write a controller in any language thanks to the simulator's server mode (i.e. one just have to write a client that connect to the simulator through a socket).
- The simulator can switch directly from simulation mode to real-world mode. This enables your controller to directly take control of a real-world robot. This is currently limited to Khepera robots but may be extended.
- The simulator can be easily interfaced with EO (the Evolutionary library), enabling the evolution of robot controllers
- The simulator currently works under Linux, Windows and Mac OS X.

For further information and download, please refer to : http://www.lri.fr/~mary/WoB/.

5.4. Simbad Autonomous and Evolutionary Robotics Simulator

Keywords: *Java*, *evolutionary robotics*, *robot simulation*.

Participant: Nicolas Bredèche [correspondent].

Abstract: Simbad is an open source Java 3D robot simulator for scientific and educational purposes (Authors: Louis Hugues and Nicolas Bredèche). Simbad embeds two stand-alone additional packages: (1) a Neural Network library (PicoNode) and (2) an Artificial Evolution Engine (PicoEvo). The Simbad package is targeted towards Autonomous Robotics and Evolutionary Robotics for research and education. The packages may be combined or used alone. In the scope of Research in Evolutionary Robotics, the Simbad package package helps quick development of new approaches and algorithms thanks to the complete and easy-to-extend libraries. Real-world interface can be easily written to transfer simbad controllers to real robots (the Khepera interface is available). The open source nature of the project combined with easy-to-understand code makes it also a good choice for teaching Autonomous and Evolutionary Robotics. Simbad is used in several AI and robotics courses: IFIPS engineering school (4th and 5th year) ; Master 1 at Université Paris-Sud ; Modex at Ecole Polytechnique.
Please refer to : http://simbad.sourceforge.net/.

5.5. PuppetMaster - Generic 3D Robotics Simulator

Keywords: *physics-based engine*, *robot simulation*.

Participant: Alexandre Devert [correspondant].

Abstract:

PuppetMaster is an open source C++ 3d robotic simulation framework for scientific and educational purposes. It allows to describe simulation scenarios, robot morphologies and behaviors as a C++ plugin. A visualizer makes it possible to see a plugin in action. The simulation is based on realistic physical simulations, so the range of the representable robots and simulations scenarii covers all the practical cases. It allows rapid prototyping of both control algorithm and robot morphology. Combined with a numerical optimization framework, it allows fully automatic design of robots, with simulations scenarios as fitness measure. PuppetMaster was used to design a robot control algorithm independent from the morphology, allowing tests on snake-like and multi-legged robots.

5.6. Django

Keywords: *Fast theta-subsumption*.

Participants: Jérôme Maloberti [correspondent], Michèle Sebag.

Abstract: Django is an algorithm of theta-subsumption of Datalog clauses, written in C by Jerome Maloberti and freely available under the GNU Public License. This algorithm is an exact one, with a gain of two or three orders of magnitude in computational cost over other theta-subsumption algorithms. Django uses Constraint Satisfaction techniques such as Arc-Consistency, Forward-Checking and M.A.C. (Maintaining Arc-Consistency) and heuristics based on the First Fail Principle.

Django has been widely used and cited in the literature (coll. with the Yokohama University, Japan, U. of Tufts in Arizona, USA, U. of Bari, Italy).

http://tao.lri.fr/TikiWiki/tiki-index.php?page=Django/.

5.7. OpenDP

Keywords: *Learning*, *Object-oriented*, *Stochastic Dynamic Programming*.

Participants: Olivier Teytaud [correspondent], Sylvain Gelly, Jérémie Mary.

Abstract: OpenDP is an open source code for stochastic dynamic programming, based upon the use of (i) time-decomposition as in standard dynamic programming (ii) learning (iii) derivative-free optimization. Its modular design was meant to easily integrate existing source codes: OpenBeagle (with the help of Christian Gagné), EO (with the help of Damien Tessier), CoinDFO, Opt++, and many others, for optimization; the Torch library and the Weka library and some others for learning. It also includes various derandomized algorithms (for robust optimization and sampling); other algorithms (e.g. time-pca and robotic-mapping) are underway.
OpenDP has been parallelized, and experimented on a large set of benchmark problems (included in the environment), allowing for an extensive comparison of function-values approximators and derivative-free optimization algorithms with a tiny number of iterates [21].
The merit of the OpenDP platform is twofold. On one hand, while many of the above algorithms are well-known, their use in a dynamic programming framework is new. On the other hand, such a systematic comparison of these algorithms on general benchmarks did not exist in the literature of stochastic dynamic programming, where many papers only consider one learning method, not necessarily in the same conditions than other published results. These thorough experimentations inspired some theoretical work in progress about the criteria for learning in dynamic environments, noting that cross-validation is neither satisfactory (for example the σ^2 parameter in Gaussian SVM chosen by cross-validation is usually too small in the context of dynamic programming) nor fast enough in that framework.
OpenDP has been presented at the Machine Learning Open Source Software Workshop at NIPS 2006 [25] and is also used in the Sequel team.

See main page at http://opendp.sourceforge.net/.

6. New Results

6.1. Fundamentals of Machine Learning, Knowledge Extraction and Data Mining

Keywords: *AUC-based Learning*, *Bounded Relational Reasoning*, *Constraint Satisfaction and Phase Transition*, *Feature Selection*, *Human Computer Interaction and Visual Data Mining*, *Inductive Logic Programming*, *Meta-learning and Competence Maps*, *Methodological aspects*, *Phase Transitions*.

Participants: Nicolas Baskiotis, Nicolas Bredèche, Antoine Cornuéjols, Sylvain Gelly, Michèle Sebag, Olivier Teytaud.

Abstract: This theme focuses on machine learning, knowledge discovery and data mining (ML/KDD/DM), investigating: i) the learning criteria; ii) the selection of features and hypotheses; iii) the randomized and quasi-randomized selection of examples; iv) the specificities of relational learning, in relation with phase transitions; v) the Multi-Armed Bandit framework.

Two book chapters (in French) have been published on these topics, about the fundamental statistical elements of machine learning [9] and the challenges of data mining [13].

Many activities below will refer to the PASCAL (Pattern Analysis, Statistical Modelling and Computational Learning) Network of Excellence (http://www.pascal-network.org), 2003-2007, which involves most major research groups in ML in Europe, including TAO and SEQUEL from INRIA Futurs. M. Sebag, in charge of the Université Paris-Sud site in Pascal, is manager of the Pascal Challenge programme and member of the Pascal Steering Committee.

6.1.1. New learning criteria

Non-convex criteria.
The flexible and effective evolutionary optimization (EC) framework enables us to explore new and non-convex learning criteria. The combinatorial Area Under the ROC Curve (AUC criterion) is optimized by the *ROC-based GEnetic learneR* (ROGER) algorithm, which has been successfully applied to bio-informatics and text mining, to ranking candidate terms during the terminology extraction step[1]. ROGER has been extended to other criteria, motivated by low quality datasets in bioinformatics and inspired by the Energy-based learning framework proposed by Y. Le Cun (2006); these criteria are investigated by A. Rimmel (PhD student under A. Cornuéjols' and M. Sebag's supervision). Along the same line, the search for stable patterns in spatio-temporal data mining was formalized as a multi-objective multi-modal optimization problem, applied to functional brain imaging[2] (ACI NIM NeuroDyne contract, in coll. with Hôpital La Pitié Salpétrière, LENA). In this perspective, the discriminant power of a hypothesis/pattern is handled as one among the objectives [33] (invited Keynote speech at COGIS 2006).

Interestingly, EC can most naturally be used to generate many hypotheses (each run provides a new hypothesis, conditionally independent of the others given the dataset). These hypotheses can be plugged for free in an ensemble learning setting, with significant gains in terms of accuracy.

This line of research differs from the mainstream ML, mostly considering convex criteria for the sake of solution unicity and optimization feasibility. Interestingly, while recent advances in ML have been concerned with integrating the AUC criterion in the convex optimization setting through a quadratic number of constraints (Joachims 2005), one switches to greedy heuristics in order to keep the computational cost under acceptable limits.

Criteria and bounds

[1] Preference Learning in Terminology Extraction: A ROC-based approach. J. Aze, M. Roche, Y. Kodratoff, M. Sebag In Applied Stochastic Models and Data Analysis (AMSDA), 2005.

[2] A Multi-Objective Multi-Modal Optimization Approach for Mining Stable Spatio-Temporal Patterns. M. Sebag and N. Tarrisson and O. Teytaud and S. Baillet and J. Lefevre In Proc. Int. Conf. on Artificial Intelligence, IJCAI 2005, L. Kaelbling Ed, 2005. IOS Press pp 859-864

Learning Bayesian Networks (BN) mixes non-parametric and parametric learning, as one must identify the structure of the network together with its weights (conditional dependency tables). S. Gelly and O. Teytaud have proposed a new complexity measure, accounting for the non-parametric complexity besides the standard number of weights [7].

Furthermore, a loss-based criterion has been proposed for the parametric learning task. While this criterion is more computationally demanding, it is more stable than the standard one, and should be preferred in particular when dealing with small datasets.

Empirical results demonstrate substantial improvements compared to the state of the art, even in the limit of large samples. Other results, combining the above with classical learning theory, include proofs of convergence to a minimal sufficient structure.

6.1.2. Selection of Features and Patterns

Feature Selection arises as a pre-processing selection task for ML, while Pattern/Hypothesis Selection is viewed as a post-processing selection task in ML or DM.

Actually, irrelevant features severely hinder the learning task, in terms of computational cost as well as predictive accuracy, particularly when the number of examples is small and/or when the signal to noise ratio is low, as is the case for micro-array analysis. Two methods inspired from ensemble methods were proposed for feature selection in 2005. The theoretical study of these approaches is tackled by Romaric Gaudel (PhD student under M. Sebag and A. Cornuéjols's supervision).

Theoretical studies about feature and pattern selection have been pursued, in relation with the PASCAL Network of Excellence. The general problem of type I and type II errors for simultaneous hypothesis testing (respectively corresponding to the selection of irrelevant hypotheses and the pruning of relevant ones) is thoroughly investigated. One conference paper [35] and one book chapter [34] have been accepted for publication about the quality-measures and statistical validation of pattern-extraction. They involve bootstrap estimates for non-independent-rules-selection as in the case of rule-extraction and surveys of standard measures of quality.

A Pascal Theoretical Challenge was launched by O. Teytaud et al. (http://www.lri.fr/~teytaud/risq/). The Challenge workshop is scheduled May 14-15th, Paris.

6.1.3. Resampling algorithms

Resampling (e.g., bootstrap) is a well-known stochastic technique for building more robust estimates. Basically, the idea is to use several subsamples of the whole sample set in order to i) estimate confidence intervals; ii) reducing the learning bias; or iii) reducing the computational cost. In practice however, resampling must achieve some tradeoff between the achieved stability improvement and the overall computational cost.

An original approach was proposed, based on quasi-random resampling and inspired from low-discrepancy sequences [47]. While quasi-random sequences are commonly used e.g. in $[0,1]^d$ and can be defined for various continuous distributions (through the use of copula functions), they are not straightforward in discrete spaces. The goal is to build M subsamples of a sample of size N, such as they are more uniformly distributed than if independently uniformly drawn. The proposed approach is based on rewriting bootstrap laws using multinomial laws and cumulative-distribution-functions. The generality of the approach is demonstrated as it applies to cross-validation, BSFD (a data-mining algorithm for simultaneous-hypothesis-testing), and bagging (ensemble methods for learning), with stability improvements.

6.1.4. Relational learning and phase transitions

Relational Learning, a.k.a. Inductive Logic Programming (ILP) is concerned with learning from relational examples such as chemical molecules (graphs), XML data (trees), and/or learning structured hypotheses such as toxicological patterns (graphs, sequences) or dimensional differential equations (mechanical models).

One additional difficulty of learning in structured domains is that the covering test checking whether a given hypothesis covers an example (theta-subsumption), is equivalent to a NP hard constraint satisfaction problem (CSP). A most efficient theta-subsumption algorithm, Django (still the best) based on the reformulation of theta-subsumption as a binary CSP and using specific datastructures, has been devised by Jérôme Maloberti (see Section 5.6).

As for CSPs, the worst-case complexity framework proves to be exceedingly pessimistic and useless for ILP. For this reason, the statistical complexity framework based on the use of order parameters, first developed for CSP and referred to as *Phase transition* paradigm, has been transported to ILP (coll. L. Saitta and A. Giordana, U. Piemonte, Italie), with many important results about the scalability of ILP. This work has been extended to the grammatical inference framework in 2005[3].

On-going work by Raymond Ros (PhD student under A. Cornuéjols and M. Sebag's supervision) resumes the above study applied to the cribbling of molecules for the ACCAMBA IMPBIO ACI, exploring the representation of molecules as SMILE sequences.
The so-called phase transition paradigm relies on the definition of order parameters (e.g. tightness and hardness of the constraints, size of clauses and examples in relational learning, alphabet size and number of states in finite state automata), and studies the empirical behavior of the algorithm at hand through extensive experimentations on random problems built uniformly from the order parameters. The result of such studies can most conveniently be summarized through the algorithm *Competence Map*; these competence maps in turn provide a principled way for selecting the algorithm most appropriate on average conditionally to the position of the problem at hand. On-going experiments aim at the competence maps of learning algorithms as function value estimators in the OpenDP framework (see section 5.7).

It must be emphasized that this approach significantly differs from an analytical algorithmic study; instead, it postulates that many heuristics are packed into really efficient algorithms, the interaction of which is hardly amenable to analytical modeling. Therefore, an empirical framework originating from natural and physical sciences is a useful tool to determine the regions in the problem space where an algorithm generally fails or succeeds.

6.1.5. Exploration vs Exploitation and Multi-armed Bandits

Many problems can be cast as an Exploration *vs* Exploitation dilemma, where one wants to both identify some best action (and must thus explore the set of actions) and maximize its current reward (and wants thus play the best action identified so far).

The maximization of the cumulated reward, referred to as Multi-Armed Bandit problem, has been intensively studied in Game Theory and Machine Learning; an optimal algorithm dubbed UCB (Upper Confidence Bound) has been proposed by (Auer et al. 2002). Its extension to tree-structured options, referred to as UCT, has been proposed by (Kocsys and Szepesvari 2006).

During his internship, Yzao Wang jointly supervised by S. Gelly and O. Teytaud, and R. Munos and P.A. Coquelin from the Center for Applied Maths in Ecole Polytechnique, has built a computer Go program based on UCT, named MoGo. MoGo has been extremely successful: it won the last four tournaments in KGS-computer-Go (http://www.weddslist.com/kgs/past/index.html) and it is ranked first among 142 programs in the championship since August 2006 (http://cgos.boardspace.net/9x9.html).
MoGo was presented at the Demos session at NIPS 2006, with an oral presentation at the Online Trading of Exploration and Exploitation NIPS Workshop 2006 [45], [42]. It must be emphasized that the game of Go has replaced the game of Chess as touchstone of modern AI; the extreme difficulty of Go is due to i/ the lack of a reliable evaluation function; ii/ a huge branching factor.

[3] Phase Transitions within Grammatical Inference. N. Pernot and A. Cornuéjols and M. Sebag In Proc. Int. Conf. on Artificial Intelligence, IJCAI 2005, L. Kaelbling Ed, 2005. IOS Press pp 811-816

The Exploration *vs* Exploitation dilemma has also been studied with respect to fast dynamic environments, motivated by News Recommendations. This application was explored as a Challenge of the Pascal Network of Excellence proposed by the Clarity Touch Company (http://www.pascal-network.org/Challenges/EEC/). The Adapt-EVE algorithm proposed by C. Hartland, S. Gelly (both are PhD student under N. Bredèche and M. Sebag's supervision), N. Baskiotis (PhD student under M. Sebag's supervision), O. Teytaud and M. Sebag won the prize from the Clarity Touch Company and was presented at the Online Trading of Exploration and Exploitation NIPS Workshop 2006 [46]. Adapt-EVE combines UCB with i) a standard change-point-detection test based on Page-Hinkley statistics; ii) a transient strategy, referred to as Meta-Bandit, handling the sequel of a change-point detection; iii) a discount strategy, allowing for more forgetful bandits.
Last, the Exploration *vs* Exploitation dilemma was considered in the framework of Statistical Software Testing. A previous approach, pioneered by A. Denise, M.-C. Gaudel and S. Gouraud, built test sets by uniformly sampling the paths in the control flow graph of the program. The limitation is that, for large programs, a huge percentage of program paths are infeasible (no input values would lead to exert the path). A generative approach, iteratively exploiting and updating some distribution on the program paths, was proposed by N. Baskiotis and M. Sebag; the gain is about two orders of magnitude compared to the previous approach[4] (invited Keynote speech at *Learning Dialogue*, Barcelona 2006).

6.2. Fundamentals of Evolutionary Computation

Keywords: *Asymptotic convergence rate*, *Convergence of evolutionary algorithms*, *Estimation of Distribution*, *Evolution Strategies*, *Self-adaptivity*.

Participants: Anne Auger, Nicolas Bredeche, Alexandre Devert, Sylvain Gelly, Mohamed Jebalia, Marc Schoenauer, Michèle Sebag, Olivier Teytaud.

Abstract: Evolutionary Computation (EC) is a unifying framework for population-based optimization algorithms. It relies on a crude imitation of the Darwinian evolution paradigm: adapted species emerge because of the interplay between natural selection and blind variations. Evolutionary algorithm design starts with the choice of a representation (i.e. the choice of the search space to explore), of the corresponding variation operators (crossover, mutation), the crafting of the fitness function, and the tuning of the many hyper-parameters (in particular, those related to the way "natural" selection is performed). In that respect, historical approaches mainly differ by the search space they work on: genetic algorithms work on bit-strings, evolution strategies on real-valued parameters, and genetic programming on structured programs – even though some significant differences also exist in the way selection is applied.

EC is now widely acknowledged as a powerful optimization framework dedicated to ill-posed optimization problems. The main reason for its efficiency comes from its flexibility to incorporate background knowledge about the application domain into the representation and the variation operators, as well as algorithmic procedures from other areas into its own variation and selection routines. A quick introduction to the field, in French, can be found in the "evolutionary" chapter of the teaching material from Grégoire Allaire's Ecole Polytechnique course on Structural Design *Conception optimale des Structures*, recently published [12]. EC for Numerical Engineering is also presented and discussed as a chapter in *Modélisation Numérique: défis et perspectives, Hermès 2006* [10].

6.2.1. Convergence Analysis for Evolution Strategies

Almost all stochastic algorithms in Computer Science are implemented using pseudo-random sequences to simulate random distributions. However, quasi-random sequences sometimes offer better properties in term of uniformity criteria. [22] addresses the case of quasi-random mutations in Evolution Strategies, and shows that all quantiles of standard estimates of the off-line result of the algorithm (i.e. both lucky and un-lucky runs) are improved by derandomization. Various quasi-random mutations are proposed, and some of them can be easily applied to many variants of Evolution Strategies (e.g. (1,λ)-ES, (μ/μ,λ)-ES, (μ,λ)-ES with λ large enough) with significant improvements in dimensionality 1-50. In particular, this generality (*all* quantiles are improved) shows that no robustness argument applies against derandomization.

[4] A Machine Learning Approach For Statistical Software Testing, Nicolas Baskiotis, Michèle Sebag, Marie-Claude Gaudel, Sandrine-Dominique Gouraud, 20th International Joint Conference on Artificial Intelligence, 2007, to appear.

Another paper[39], generalizes previously known lower bounds to any comparison-based algorithm (including e.g. direct search methods, and not only evolutionary methods). The theorem, using entropy numbers of the domain, has very weak assumptions and matches existing upper bounds. The obtained lower bound holds for any

We also studied [23] the idea of defining rigorously frameworks in which the optimal algorithm exists. First, this paper shows the optimality of comparison-based algorithms in the robust-framework (worst case on increasing transformations of the objective function). This result is the counterpart of the previous paper: comparison-based algorithms are slower, but optimal for some robustness-criterion. Also with the idea of using optimality in optimization, the paper shows that under some prior distribution of fitness and with a maximum (a priori known) number of objective-function-evaluations, one can define the notion of "optimal optimization algorithms", and implement such algorithms. The "optimal algorithm" is computationally very expensive, but approximations are proposed and are relevant for the framework of EDA (estimation of distribution algorithms).

Similar ideas have been applied to study the complexity of approximating Pareto-sets with large number of conflicting objectives in the context of Evolutionary Multi-Objective Optimization. Strong lower bounds have been derived, mainly leading to the conclusion that multi-objective problems with many conflicting objectives can not be fully solved off-line [40]. This is in accordance with practice, as practitioners admit that they can not deal in a fully off-line manner with many conflicting objectives, and suggests to move to on-line interactive methods when the number of conflicting objectives is large. Note that the paper also provides theoretical foundations for the use of methods based on the removal of moderately conflicting objectives.

6.2.2. Genetic Programming

Genetic Programming (GP) is a technique to evolve programs, represented as parse-trees. GP can directly be used for supervised learning, in which case the output of the program (the tree) is the class a given example belongs to. In this framework, in the continuation of C. Gagné's PhD thesis and in collaboration with M. Tomassini (U. Lausanne) and M. Parizeaux (U. Laval à Québec), we studied the influence of several heuristics for the choice of the best GP classifier after a multi-objective GP run on the bloat (the uncontrolled growth of the sizes of the trees) [19].

But GP can also be used for Machine Learning in a more indirect way: the choice of the kernel is known to be a major issue when using a kernel-based learner. In [20], GP is used to build the optimal kernel for a given learning task; Furthermore, the complexity of the algorithm is reduced by co-evolving the subset of samples and the kernels, to avoid learning on all examples at once.

Another use of GP is proposed in Alexandre Devert's PhD work, namely an embryogenic approach to the Optimum Design problem. This work is described in more details in section 6.3.3.
On the theoretical side, the work using hints from learning theory for symbolic regression and presented at CAp 2005 has been continued: This work uses methods from statistical learning theory to prove that bloat cannot be avoided in various standard frameworks and to propose new method to fight bloat that are proved to work, and experimentally validated by some extensive experiments. Those extended results have been presented at Dagstuhl seminar[44], and published in a journal version (in French) [6], while an English version has been submitted to an international journal.

Related works [41] include some complexity and computability results showing in some minimax sense the not-too-far-from-optimality nature of simulation-based and selection-based methods like genetic programming for mining spaces of Turing-Computable functions.

6.2.3. Surrogate models in Evolution Strategies

The work about surrogate models, that was started in the Tao team by K. Abboud in his Ph.D. thesis, defended in 2004, has been continued by the work of Y. Bonnemay (co-supervised between Saint-Gobain and Tao) and O. Teytaud. This work concern the theoretical properties of surrogate models, and an empirical study on (i) standard test difficult test functions (ii) industrial problems. Also, Tao is leading the workpackage *General meta-models* within the RNTL project OMD (*Optimization Multi-Disciplinaire*) (see section 6.4.5).

We compared surrogate models and Estimation of Distribution Algorithms in [24] in the particular framework of very small numbers of iterates (expensive fitness-functions); more information about this is provided in section 6.2.4. We also developed a mathematical analysis of very different ways of including learning in optimization; a draft of this work, based on Bellman's decomposition, is available at http://www.lri.fr/~teytaud/optim.pdf.

6.2.4. Estimation of Distribution Algorithms

Estimation of Distribution Algorithms (EDAs) proceed by alternatively sampling and updating a distribution on the search space. The sampled individuals are evaluated, i.e. their fitness is computed, and the distribution is updated and biased toward the best individuals in the current sample. Extensions of this framework to continuous optimization was initialized by Ducoulombier & Sebag (1998)[5].

On-going work (Ph.D. Celso Ishida, co-advised with A. Pozo, Universidad Federale do Parana, Brazil) is concerned with using mixtures of distributions, borrowing to the MIXMOD EM-like approaches developed in the SELECT project at INRIA, to extend EDAs to multi-modal optimization.
On the theoretical side, we studied a particular class of EDA [24] in the particular hot framework of expensive optimization functions (which is included in the OMD RNTL presented in section 6.4.5, as well as in the RedOpt working group – http://norma.mas.ecp.fr/wikimas/RedOpt). This paper provides conservative upper bounds that provide hints about parametrization of EDA, in particular depending on the available ressources (number of function-evaluations). In spite of the fact that the analysis is based on very conservative tools from VC-theory, the resulting algorithm is efficient for very frugal frameworks in which the number of function-evaluations is very moderate. The theoretical analysis emphasizes the importance of the dependency of the algorithm on the population size, that should be chosen much larger when the number of iterates is larger, and should also be much larger than usually done when robustness is a main goal.

6.3. Modelling and Control of Complex Systems

Participants: Nicolas Bredèche, Antoine Cornuéjols, Alexandre Devert, Mary Felkin, Sylvain Gelly, Cédric Hartland, Jérémie Mary, Miguel Nicolau, Raymond Ros, Marc Schoenauer, Michèle Sebag.

Abstract: Several research directions were initially targeted toward Robotics, but are in fact relevant to more general Complex Systems. This includes, beside Robotics, activities in e-Science and Grid Modelling, but also Evolutionary Design and Inverse Problems, that have been part of Tao activities for many years, as well as new studies related to Complex Networks.

6.3.1. Robotics

Four directions have been explored: the first two ones are related to the knowledge transfer from human experts to the robot controller; the third one, investigating the reality gap, aims at transforming an *in-silico* optimized controller into an *in-situ* competent one. The last one is concerned with optimization of locomotion for several robot morphologies using Central Pattern Generators.

Imitation Learning
In collaboration with Lutins (U. Paris 8 - Cognitive Psychology group, Herobot contract), we have considered the problem of imitation learning with visual validation. Specifically, a robot controller aimed at reproducing the behavior of human subjects with impaired perception (exploring a maze in order to find some object), has been devised using the subsumption architecture. *In silico*, this controller experimentally displayed a plausible behavior, providing the psychologists with relevant insights and suggesting fruitful exploration heuristics. Experiments *in situ* are under way [6].

[5] Extending Population-Based Incremental Learning to Continuous Search Spaces, in Th. Bäck et al., Eds, PPSN'98, LNCS 1498, pp 418–427, Springer-Verlag, 1998

[6] C. Tijus, N. Bredeche, Y. Kodratoff, M. Felkin, C. Hartland, E. Zibetti, V. Besson. Human Heuristics for a Team of Mobile Robots. 5th IEEE International Conference on Research, Innovation and Vision for the Future (RIVF'07).

Fuzzy controllers and prior knowledge
Among the representations investigated for robot control are neural nets and fuzzy controllers. Carlos Kavka, from San Luis University (Argentina), working under Marc Schoenauer's supervision, extended the standard evolutionary design of fuzzy systems to Voronoi diagram-based representation (see section 6.3.3). This representation does not only allow for more flexible decision boundaries (whereas standard approaches consider hyper-rectangles); it also enables the easy embedding of expert rules (e.g. "Go forward if there is no obstacle ahead") and their automatic optimization, specialization or generalization, depending on the application problem. This approach demonstrated its efficiency for evolutionary robotics[7] and was extended to recurrent fuzzy systems, endowing the robot with some self-managed memory. A comprehensive description will be found in Carlos Kavka's PhD [1], defended in July 2006 at Université Paris-Sud.

Anticipation and the Reality Gap
Earlier work explored the use of an anticipation module in order to enhance the autonomous controller stability[8]. This anticipation module was reconsidered to facilitate the transfer *in situ* of a robotic controller after its *in silico* optimization. Specifically, the anticipation module trained *in silico* provided the controller with an additional information, the residue (difference between the predicted and actual state in the next time step). This residue was exploited for the on-line adaptation of the controller; experiments successfully demonstrate the on-line robot recovery under motor perturbations [29].

Locomotion Optimization
In the context of locomotion, we addressed the problem of locomotion for snake and legged robots. These kinds of locomotion share the fact that they rely on oscillatory signals to generate appropriate locomotion patterns. Central Pattern Generator (CPG), inspired from biology, are Dynamical Systems that can be interconnected and are characterized by limit cycles that can generate a relevant activity pattern for locomotion. Moreover, convergence towards this very limit cycle makes is very useful when it comes to recovering from punctual control errors (sliding, hardware temporary failure, etc.). In this scope, we have studied a well known implementation of a CPG and optimized parameters for several topologies corresponding to several robot morphologies. Our approach provides an efficient way to automatically learn locomotion independently of the morphology - Experiments were conducted using the *PuppetMaster* simulator (section 5.5) with a snake-like robot and a hexapodal robot in realistic physics-based simulation (Enguerran Colson, Master 2 recherche, 2006).

6.3.2. e-Science and Grid Modelling

As Cecile Germain joined the TAO group in 2005, her strong expertise in grid computing opens new and strategic perspectives along several main directions.

[7] C. Kavka and M. Schoenauer. Evolution of Voronoi-based Fuzzy Controllers. In Xin Yao et al., eds, PPSN'04, LNCS 3242, Springer Verlag, 2004.

[8] Robea contract, coll. LIMSI. N. Godzik and M. Schoenauer and M. Sebag, Robustness in the long run: Auto-teaching vs Anticipation in Evolutionary Robotics. In X. Yao et al., Eds, *Proc. PPSN VIII*, pp 932-941, LNCS 3242, Springer Verlag, 2004

A first direction, explored in the Programme Pluri-Formation DEMAIN (*Des DonnéEs MAssives Aux InterpretatioNs*, starting Dec. 2006, headed by C. Germain), is that of e-Science (http://www.lri.fr/ cecile/DEMAIN/DemainSc.htm). With LRI, LAL (Laboratoire de l'Accélérateur Linéaire), Lab. Maths, IBBMC (Institut de Biochimie et Biophysique Moléculaire et Cellulaire) and Supelec as main partners, DEMAIN aims at developing pump-priming projects on the Orsay campus, concerned with the principled exploitation of the datasets gathered in LAL and IBBMC, using advanced algorithms in machine learning, data mining and optimization. A typical application concerns the analysis of the Auger experiment in collaboration with A. Cordier (LAL). This junction between the LAL and the Tao group was instrumental in recruiting Balazs Kegl as CNRS CR1, researcher in machine learning at LAL and correspondent of the Tao group. DEMAIN avails the computing facilities gathered by the LAL, namely the EGEE (Enabling Grids for e-Science in Europe) grid. Cecile Germain currently chairs the Short Deadline Jobs (http://egee-intranet.web.cern.ch/egee-intranet/NA1/TCG/wgs/sdj.htm) working group in the EGEE Network of Excellence.
A second and more daring research direction is that of Grid Modelling. A complex system, the grid can hardly be modeled through a-priori analysis: its topology and state at any time can only be estimated; the grid usage, based on a mutualisation paradigm, reflects the collective behavior of the users and results in an uncontrolled and unforeseeable load on the system. Interestingly, the modelling of the computing system is viewed as the first step (building self-aware systems) in the Autonomic Computing effort, declared as a top priority for the IBM company since 2001 (http://www.research.ibm.com/autonomic/overview/). First steps toward the EGEE control have been done in 2005 [9]. A Pascal Challenge related to Grid Modelling has been accepted in Dec. 2006 (coll. TAO, LAL, Technion). Xiangliang Zhang (PhD student under M. Sebag and C. Germain's supervision) recently started the modelling of the EGEE grid.
The Grid Modelling Challenge is also relevant to the KD-Ubiq Coordination Action, started in 2006 (http://www.kdubiq.org/). M. Sebag is responsible for the Work Package Benchmarking.

Lastly, Cécile Germain chairs the ACI MD AGIR (http://www.aci-agir.org) contract (starting sept. 2004), concerned with medical data mining and more precisely medical imaging through grid computing [10]. A multi-disciplinary project, AGIR gathers researchers in computer science, physics and medicine from CNRS, Université Paris-Sud, INRIA, INSERM and hospitals.
Julien Perez (PhD student under C. Germain and A. Osorio from LIMSI supervision) is concerned with the reconstruction of 3D images through mining the logs of the PTM3D software, developed at LIMSI and ported on the grid[11]. This study ultimately aims at grid-aware mining algorithms.

6.3.3. Evolutionary design

Earlier work about evolutionary design, applied to Topological Optimum Design of Mechanical Structures[12] or Architecture (EZCT contract[13]) explored several shape representations overcoming the limitations of the standard bitarray representations, including Voronoi diagrams (see also Section 6.3.1).

Another representation, that of construction plans, was investigated by A. Devert (PhD under N. Bredèche' and M. Schoenauer's supervision). This representation is close to the so-called "embryogenic" representations, in which the evolution optimizes a program that actually builds the phenotype (here, the actual structure). The program, a directed acyclic graph, describes a sequence of possible actions (drop, move in straight line, rotate, ...). Such representation addresses some deep requirements for design, e.g. modularity, re-usability, and not least in the domain of Structural Design, constructibility, as only feasible moves are planned in the program.

[9] C. Germain and D. Monnier-Ragaigne. Grid Result Checking. In Procs. 2nd Computing Frontiers, Ischia, Mai 2005.
[10] C. Germain, V. Breton, P. Clarysse, Y. Gaudeau, T. Glatard, E. Jeannot, Y. Legré, C. Loomis, J. Montagnat, J-M Moureaux, A. Osorio, X. Pennec et R. Texier. Grid-enabling medical image analysis, Journal of Clinical Monitoring and Computing, 19(4-5), 339-349, 2005.
[11] PTM3D has been part of the first EGEE review and of the HealthGrid demonstrations at SC'05
[12] H. Hamda, F. Jouve, E. Lutton, M. Schoenauer and M. Sebag. Compact Unstructured Representations in Evolutionary Topological Optimum Design. Applied Intelligence, 16, pp 139-155, 2002.
[13] Results of chair designs have been exposed in the Innovative Design Techniques section of the ArchiLab exhibition in Orléans in 2005, and have been acquired by the Beaubourg Modern Art Museum.

The scalability of the approach (handling up to a few hundred modules) is higher by an order of magnitude than that of existing approaches [17], [18].

On-going work extends the construction plan representation towards embryogenesis. Specifically, the goal is to both optimize the "basic cell", and the connectivity of a network made of some hundred basic cells under locality and network diameter constraints (see also section 6.3.5).
Another daring design problem concerns the optimization of mesh topologies (Airbus Contract), motivated by the fact that mesh topology design requires both considerable time and expertise from the designers; an extensive corporate knowledge is encapsulated in the mesh topology archive. The Airbus project involves three phases; the first one is about representing available meshes in a tractable way, using propositionalization (the intrinsic description of a mesh is through a few thousands/ hundred thousands of finite elements); the second phase is concerned with characterizing good meshes (using a one-class learning approach, as only good topologies are stored); the third phase uses the above characterization to derive new good meshes. Due to the unexpected leave of Mathieu Pierres, PhD in May 2006, this project has not made any progress in 2006. Damien Tessier will take over the project in 2007.

6.3.4. Inverse problems

Inverse Problems (IP) aim at determining unknown causes based on the observation of their effects. In contrast, direct problems are concerned with computing the effects of (exhaustively described) causes. Inverse problems are often mathematically ill-posed in the sense that the existence, uniqueness and stability of solutions cannot be assured.

IPs are present in many areas of science and engineering, such as mechanical engineering, meteorology, heat transfer, electromagnetism, material science, etc. The TAO project has focused on the problems of system identification, modeling physical (mechanical, chemical, biological, etc.) phenomena from available observations and current theories.

Domain Knowledge for Seismic Velocity Identification

A long collaboration with IFP, the seismic inverse problem aims at identifying some underground characteristics from recorded seismic data. Earlier work[14] has been using Voronoi diagrams for representing the underground, demonstrating that the available objective function was not sufficient to enforce plausible solutions (e.g., some underground profiles with not so bad fitness were geophysically absurd).

Vijay Pratap Singh (PhD under Marc Schoenauer's supervision) remedied the above limitations through a more knowledgeable representation, evolving an initial state of layered underground as well as the geological conditions across the geological ages. However, though it allowed good results on the geological problem[15], and also let to a patent in 2005, this representation didn't allow to successfully solve the geophysical problem. The only way to handle the large computational cost of evolutionary methods was to introduce domain knowledge wherever it was possible. But the results were a computational gains of orders of magnitude! These results are extensively described in his PhD, defended in December 2006 [2].

Representations for isotherm law in chromatography In the framework of the ACI NIM *Nouvelles Interfaces des Mathématiques*, Marc Schoenauer is part of the *Chromalgema* project, whose aim is the identification of the isotherm function in analytical chromatography. This is an inverse problem for which the direct problem is solved by standard numerical approaches (e.g. Godunov scheme for Non-Linear Hyperbolic Systems).

[14] F. Mansanné, Analyse d'Algorithmes d'Évolution Artificielle appliqués au domaine pétrolier : Inversion sismique et approximation de fonctions, PhD Université de Pau, 2000

[15] V.P. Singh, M. Schoenauer and M. Léger, A geologically-sound representation for evolutionary multi-objective subsurface identification, in B. McKay et al., Eds, CEC'05, pp 454-462, IEEE Press, 2005

When the unknown isotherm function is sought as a rational fraction of the concentrations (e.g. in the family of so-called "Langmuir" models), the inverse problem amounts to parametric optimization. A recent improvement was to use the recent "CMA-ES" method and its refinements (the restart strategy). On-going work is related to the hybridization of evolutionary and gradient methods: what is the best hybridization method: sequential (and when to switch), fine-grained (and what amount of local optimization to perform inside evolution), or both? Those results, as well as validation on real-world data, will be presented exhaustively in Mohamed Jebalia's PhD dissertation.

EvoTest
The automatic generation of test data can indeed be seen as an inverse problem: what data should be input to the program under test to reach this or that instruction (for structural testing), or to trigger such or such functional error (for functional testing)? TAO is part of the European STREP *EvoTest*, that started on October 1. 2006, funded under FP6 FET "complex systems" call. The coordinator is ITI, University of Valencia, and the main partners are Daimler-Chrysler, Berlin, Franhofer FIRST, Berlin, and Kings College, London.

6.3.5. Optimization and Identification of Complex Networks

This section describes prospective work that has started in TAO in 2006, and hence has not yet resulted in any publication.

Approximate stochastic simulation of chemically reacting system
Two mathematical models exist for describing the time behavior of chemical system: In the deterministic model, the time evolution of the chemical species is modeled as a set of ordinary differential equations; In the stochastic model, the different species are random variables obeying the chemical master equation that takes into account their inherent fluctuations and correlations, that cannot be neglected when dealing with biological systems where the overall number of molecules is usually small. An exact simulation algorithm for the chemical master equation was introduced by Gillespie in 1977 [16].

Anne Auger, who joined TAO in 2006 as Chargée de Recherche, recently proposed (with co-authors) a new accelerated scheme [3] where the complexity is reduced in the case of (moderately) stiff systems.

Optimizing the topology of large neural networks
The performance of large networks of small computational units like neural networks or ...the next generation of multi-core micro-processors highly depends on the topology of the network. Fei Jiang, that started a PhD in September, following his engineer internship, works on both the direct problem (what is the influence of the topology of given networks on their performance) and the inverse problem (how to design optimal networks for a given task). This thesis is co-directed by Hugues Berry, CR1 in the Alchemy project, and Marc Schoenauer.

Genetic Regulatory Networks
The GENNETEC European project, funded under FP6 FET "complex systems" call, has begun in October 2006, and deals with Genetic Regulatory Networks (GRN): W. Banzhaf's GRN model[17], is a generative model where an interaction network between genes emerges from a series of genetic evolutionary variations. The resulting system of ODEs can then be solved to compute the evolution of protein concentrations. The work of Miguel Nicolau, hired on Gennetec project in October 2006, will be to tune the evolutionary variations of the genome, in order to control both the topology of the interaction network and the behavior (transient as well as steady-state) of the system of proteins.

Cellular Evolutionary Design

[16] D.T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, Vol. 81, pp 2340-2361, 1977.
[17] W. Banzhaf, Artificial Regulatory Networks and Genetic Programming, in Rick L. Riolo and Bill Worzel, Eds, Genetic Programming Theory and Practice, pp 43-62, Kluwer, 2003

Recent research direction by A. Devert is also concerned with the evolution or large networks of interacting elements: to cope with the scaling issue of evolving construction plans (section 6.3.3), the idea is to evolve local rules for some "cells" that will exchange "chemicals", and the steady-state of those chemicals will describe the target structure. However, the neighborhood topology of the whole domain can be evolved, too ...

6.4. Other applications

Participants: Thomas Heitz, Yves Kodratoff, Claire Le Baron, Marc Schoenauer, Michèle Sebag, Yann Semet, Olivier Teytaud.

Abstract: The TAO group is also historically involved in other applications of either Machine Learning or Evolutionary Computation that are not directly linked to its main streams of research. They are surveyed below.

6.4.1. Text Mining

Text Mining (TM) is concerned with exploiting/transforming documents to achieve particular tasks. The difficulty lies in the delicate balance to keep between texts, transformations and tasks. Problem resolution implies the existence of cognitive entities, called concepts of specialty, necessary to the resolution of the current tasks.
A key data preparation step in Text Mining, Term Extraction selects the terms, or collocation of words, attached to specific concepts. The task of extracting relevant collocations can be achieve through a supervised learning algorithm, exploiting a few collocations manually labeled as relevant/irrelevant. In Thomas Heitz' PhD work, an evolutionary learning algorithm (ROGER, see section 6.1.1), based on the optimization of the Area under the ROC curve criterion, extracts an order on the candidate terms. The robustness of the approach was demonstrated on two real-world domain applications, considering different domains(biology and human resources) and different languages (English and French)[18]. All detailed will be available in Thomas' forthcoming PhD dissertation.

After having organized the fist edition, the team has organized the second French text-mining challenge (DEFT: DÉfi Fouille de Textes), which consisted of topics segmentation from French corpora built for the occasion. It took place in a workshop of SDN'06 conference with twenty five participants belonging to seven francophone teams. The results by the participating teams are presented in [30].

6.4.2. Large-Eddy Simulations

Large-Eddy Simulations efficiently provide the large eddies of the flow, but not the Filtered Density Functions (FDF), that is required for evaluating some interesting quantities. As the FDF is well approximated by a two-parameters family of distributions, namely the β-distributions, we therefore only have to estimate these two parameters from the large eddies. We compared in [11] (i) various sets of variables extracted from the large eddies (ii) various approaches for the estimation of these parameters from these sets of variables. We conclude to the theoretical and practical efficiency of universally consistent non-parametric learning tools like neural networks.

6.4.3. Train scheduling

Scheduling problems are a known success area for Evolutionary Algorithms (EAs). The French Railways (SNCF) were interested to find out whether they could benefit from EAs to tackle the problem of rescheduling the trains after an incident has perturbed one (or more) train(s). At the moment, they are using in their production units a commercial software (*CPlex* from Ilog), and they experience serious difficulties when several incidents occur simultaneously on large networks.

[18] Jérôme Azé, Mathieu Roche, Yves Kodratoff, Michèle Sebag. Preference Learning in Terminology Extraction: A ROC-based approach, in: "ASMDA'05", 2005.

After some significant results on full size problems[19], showing that EAs can indeed give better and faster results that *CPlex* thanks to a very specialized "scheduler", the contract was renewed for an additional year by SNCF, and ended in May 2006. Yann Semet was thus able to polish his algorithm, and to increase the performance of EAs even more thanks to some inoculation of good solutions in the initial population. This work also allowed him to start identifying the type of problems where EAs are a better alternative than *CPlex* [38]. Note also that this work was mentioned in 2 publications in *Rail et Recherche*, the SNCF internal magazine.

6.4.4. Time-dependent planning with bounded resources

A one-year contract between TAO and Thales - Land & Vision division was concerned with temporal planning with limited resources (in 2004-2005). Two approaches have been tried: coupling Evolutionary Algorithms on the global scale with Constraint Programming to solve local (hopefully small) problems on the one hand; and using Petri Networks representing partial plans.

Only the first approach succeeded, and feasibility results for the *TGV* approach have been obtained during the contact (that ended in July 2005). Collaboration continues with Pierre Savéant (Thales) and Vincent Vidal (Université de Lens) as this approach allowed to obtain breakthrough results : it is the first Pareto approach for multi-objective temporal planning problems [37]. A CIFRE PhD position on this topic, funded by Thalès, has recently been posted.

6.4.5. Multi-disciplinary multi-objective optimization

TAO team is involved in the RNTL project on Multi-disciplinary optimization coordinated by Rodolphe Leriche (Ecole des Mines de St-Etienne), for its expertise in surrogate models. This will most probably lead to work in different application domains. Independently (even though Renault is also a partner of OMD consortium), Claire Le Baron had started a CIFRE PhD in September 2005 funded by the automobile company Renault. The (ambitious) original goal of the PhD is to optimize the complete motor of a car, thus involving structural mechanics, vibration and acoustics, combustion and thermics. But the thesis is now focusing on comparing a wide range of approaches involving some trade-off between solving a faithfull but complex optimization problem (using Evolutionary Algorithms), and solving an easy but maybe not really significant optimization problem using standard numerical methods.

7. Contracts and Grants with Industry

7.1. Contracts and Grants with Industry

Keywords: *Airbus*, *Chromalgena*, *FP6 Complex Systems call*, *IFP*, *Quantum chemistry*, *RNTL*, *Renault*, *SNCF*.

Contracts managed by INRIA

- **Chimie Quantique**, CNRS Program ACI-NIM (New Interfaces of mathematics) – 2004-2007 (8 kEur), coordinator Claude Le Bris(Cermics);
  participant: Marc Schoenauer (presentation at the ACI Winter school in January 2006).
- **Chromalgena**, CNRS Program ACI NIM (New Interfaces of mathematics) – 2003-2006 (14 kEur), coordinator F.James (Université d'Orléans) (section 6.3.4);
  participant: Anne Auger, Mohamed Jebalia, Marc Schoenauer.
- **AIRBUS** – 2004-2007 (45 kEur), was the side-contract to Mathieu PIERRES's CIFRE Ph.D. (section 6.3.3);
  Participants: Marc Schoenauer and Michèle Sebag.
- **IFP** – 2003-2006 (18 kEur), side-contract to Vijay Pratap SINGH's CIFRE Ph.D. (section 6.3.4);

[19] Y. Semet and M. Schoenauer. An efficient memetic, permutation-based evolutionary algorithm for real-world train timetabling. In B. McKay et al.,Eds, Proc. CEC'05, pp 661-667, IEEE Press, 2005

Participants: Vijay Pratap Singh and Marc Schoenauer.

- **SNCF** – 2004-2006 (2×80 kEur) : research contract, Yann SEMET, expert engineer (section 6.4.3);
  Participants: Marc Schoenauer and Yann Semet.
- **ONCE-CS** – 2005-2008 (147 kEur) European *Coordinated Action* from FP6. Coordinator Jeff Johnson, Open University, UK;
  Participants: Bertrand Chardon, Marc Schoenauer.
- **OMD-RNTL** – 2005-2008 (72 kEur) Coordinator Rodolphe Leriche, Ecole des Mines de St Etienne;
  Participants: Anne Auger, Olivier Teytaud and Marc Schoenauer.
- **Renault** – 2005-2008 (45 kEur) side-contract to Claire LeBaron's CIFRE Ph.D. (section 6.4.5);
  Participants: Claire LeBaron, Marc Schoenauer.
- **EZCT** – 2005-2006 (10 kEur) side-contract to Alexandre Devert's PhD (section 6.3.3);
  Participants: Nicolas Bredèche, Alexandre Devert, Marc Schoenauer.
- **EvoTest** – 2006-2009 (231 kEur) European *Specific Targeted Research Project* from FP6. Coordinator Tanja E.J. Vos, Instituto Tecnológico de Informática, Spain;
  Participants: Marc Schoenauer.
- **GENNECTEC** – 2006-2009 (379 kEur) European *Specific Targeted Research Project* from FP6. Coordinator François Képès, Génopôle and CNRS, France;
  Participants: Miguel Nicolau, Marc Schoenauer.

Contracts managed by CNRS or Paris Sud University

- **PASCAL**, Network of Excellence, 2003-2007 (34 kE in 2005). Coordinator John Shawe-Taylor, University of Southampton. M. Sebag is manager of the Challenge Programme.
- **KD-Ubiq**, Coordinated Action, 2005-2008 (19 kE). Coordinator Michael May, Fraunhofer Institute. M. Sebag is responsible of the Benchmarking WP.
- **Neurodyne** ACI-NIM (New Interfaces of mathematics) – 2003-2006 (6 kEur). Coordinator Sylvain Baillet, LENA, Hôpital La Pitié-Salpètriere, CNRS UPR 640;
  Participants: M. Sebag, O. Teytaud, A. Cornuéjols.
- **Traffic** ACI-NIM (New Interfaces of mathematics) – 2004-2007 (17,5 kEur). Coordinator Jean-Michel Loubès, Project SELECT;
  Participants: M. Sebag, O. Teytaud.
- **AGIR** ACI Masses de Données (section 6.3.2) - 2004-2007 (260 kEur) Coordinator Cécile Germain-Renaud. Participants are from CRAN, CREATIS, INRIA-Sophia, I3S, LPC, LIMSI, LORIA, PCRI, Centre Antoine Lacassagne, CHRU Clermont-Ferrand , Tenon Hospital, Fédération de la Mutualité Parisienne;
  Participants: Cécile Germain, Michèle Sebag, Xiangliang Zhang.
- **Herobot**, TCAN CNRS – 2004-2006 (28 kEur);
  Participant and coordinator : Nicolas Bredèche.
- **Galileo**, Programme d'actions intégrés franco-italien – 2007 (4.2 kEur);
  Participant and coordinator: Antoine Cornuéjols.

8. Other Grants and Activities

8.1. International actions

8.1.1. Management positions in scientific organizations

- Marc Schoenauer, Board Member of ISGEC (International Society on Genetic and Evolutionary Algorithms) since 2000. This Society became the ACM SIGEVO (Special Interest Group in Evolutionary Computation) in 2006, but the board remained unchanged.

8.1.2. Collaborations with joint publications

- Universidad de San Luis, Argentina [1].
- Université Laval à Québec [19].

8.2. European actions

8.2.1. Management positions in scientific organizations

- Marc Schoenauer, Member of PPSN Steering Committee (Parallel Problem Solving from Nature) since 1998.
- Michèle Sebag, Member of PASCAL Steering Committee (Pattern Analysis, Statistical Modeling and Computational Learning, FP6 NoE) since 2004.

8.2.2. Working groups

- EGEE, Enabling Grids for E-SciencE : Cécile Germain-Renaud is a member, and chair for the Working Group *Short Deadline Jobs*.
- ONCE-CS, Coordinated Action, 6th Framework Program: TAO (Marc Schoenauer) is one of the main contracting nodes, responsible of WP2 - Web Portal and Services. Bertrand Chardon is paid as engineer and works on this WP.
- PASCAL, Network of Excellence, 6th Framework Program: Michèle Sebag, corresponding member for Université Paris-Sud since 2003, Manager of the Challenge Programme since 2005.

8.2.3. Collaborations with joint publications

- Université Lausanne [19], [20].

8.3. National actions

8.3.1. Organization of conferences and scientific events

- JET, Journées Évolutionnaires Trimestrielles: Marc Schoenauer organized the first editions since their creation in 1998 until 2004. Now member of the steering Committee.
- Evolution Artificielle: the international conference on Evolutionary Computation, is organized in France every second year, and has acquired a world-wide reputation not only because of the good wine and food ...Marc Schoenauer is in the organizing committee since the first edition in 1994.
- Apprenteo, gathering the researchers of the Digiteo Lab (PCRI, CEA, SupElec, LIMSI, CMAP) now RTRA, had a second meeting on March 16th, organized by Michele Sebag.

8.3.2. Management positions in scientific organizations

- National research program on knowledge management, machine learning and new technologies ACI TCAN Traitement des Connaissances, Apprentissage, Nouvelles Technologies: Michèle Sebag, member of the steering committee, 2002-2004; Antoine Cornuéjols, member of the steering committee since 2004.
- CNRS Network "Discovering and Summarizing", Réseau Thématique Pluridisciplinaire Découvrir et résumer, RTP 12: Michèle Sebag, member of the steering committee since 2002.

8.3.3. Associations

- *Evolution Artificielle* : Marc Schoenauer, founding president (1995-2004), now member of the Executive Committee.
- *AFIA, Association Française d'Intelligence Artificielle* : Marc Schoenauer, member of Executive since 1998, former president (2002-2004) ; Michèle Sebag, member of Executive since 2000, treasurer in 2003-2004, president since 2004 ; Jérémie Mary, treasurer, 2004-2006.
- *FERA, Fédération des Equipes de Recherche en Apprentissage* : Michèle Sebag, member of the Steering Committee with Stéphane Canu, Manuel Davy and Jean-Gabriel Ganascia.

8.3.4. Collaborations with joint publications

- Université de Lens [37], [36].
- Université de Saint-Etienne [16], [10].

8.4. Honors

- MoGo, developed in the Team by S. Gelly and Y. Wang, won the two October Kgs-tournaments and the two November Kgs-tournaments (http://www.weddslist.com/kgs/past/index.html) and is first-ranked on the Cgos-server (http://cgos.boardspace.net/9x9.html).
- N. Baskiotis, S. Gelly, C. Hartland, M. Sebag, O. Teytaud won the Exploration-Exploitation Challenge http://www.pascal-network.org/Challenges/EEC/Results/ from the Pascal Network of Excellence.

8.4.1. Keynote addresses

- COGnitive systems with Interactive Sensors - March 15-17 2006, Paris. Michèle Sebag, invited plenary speaker.
- Learning Dialogues, 2 - 5 October 2006, Barcelona. Michèle Sebag, invited plenary speaker.

9. Dissemination

9.1. Animation of the scientific community

9.1.1. Editorial boards

- Marc Schoenauer is Editor in Chief of MIT Press Evolutionary Computation Journal (since 2002)
- Marc Schoenauer is Associate editor of Kluwer Genetic Programming and Evolvable Machines (since its creation in 1999), of Elsevier Theoretical Computer Science - Theory of Natural Computing (TCS-C) since its creation in 2002, of Elsevier Applied Soft Computing since its creation in 2000, and has been Associate Editor of of IEEE Transactions on Evolutionary Computation (1996-2004) and of Kluwer Journal of Heuristics (1997-2003).
- Marc Schoenauer is on the Editorial Board of the book series *Natural Computing* by Springer Verlag, and *Mathématiques Appliquées* by SMAI (Springer-Verlag).

- Michèle Sebag is member of the Editorial Board of Knowledge and Information Systems (since 2003), of Machine Learning Journal (since 2001), of Genetic Programming and Evolvable Hardware (since 2000); she has been Associate Editor of of IEEE Transactions on Evolutionary Computation (1998-2004) and of Revue d'Intelligence Artificielle (2002-2005).

9.1.2. Chair in Organizing Committees

- Marc Schoenauer was Member of the Organizing Committee of the *2nd European Conference on Complex System* in Oxford, 25-29 Sept. 2006.
- Michèle Sebag was co-chair of the Second Pascal Challenge Workshop, Venice, April 2006.
- CAP'07 (Conférence Francophone d'Apprentisage) (July, 2007): Antoine Cornuéjols, co-chair (with Jean-Daniel Zucker)

9.1.3. Program Committee Member (international events)

- Nicolas Bredèche: European Conference on Genetic Programming, ICINCO Workshop on Multi-agent Robotic Systems.
- Marc Schoenauer: Genetic and Evolutionary Computation Conference, IEEE Congress on Evolutionary Computation, Parallel Problem Solving from Nature, European Conference on Genetic Programming, Evolutionary Computation for Combinatorial Optimization Problems, European Conference on Complex Systems, ...
- Michèle Sebag: PC of ICML 06, 23rd International Conference on Machine Learning, ECML-PKDD 06, 17th European Conference on Machine Learning, 10th Conference on Principle and Practice of Knowledge Discovery from Databases, IJCAI 07, 20th International Conference on Artificial Intelligence; ILP, Inductive Logic Programming, PPSN, Parallel Problem Solving from Nature, EuroGP, European Conference on Genetic Programming, GECCO, Genetic and Evolutionary Computation Conference, CEC, IEEE Congress on Evolutionary Computation, ...
- Olivier Teytaud: Approximate Dynamic Programming and Reinforcement Learning ADPRL'07

9.1.4. Program Committee Member (national events)

- CAP, Conférence d'apprentissage: Michèle Sebag since 1999; Antoine Cornuéjols, since 1999; Olivier Teytaud since 2005.
- EA, Evolution Artificielle: Marc Schoenauer and Michèle Sebag since 1994.
- EGC, Extraction et Gestion des Connaissances: M. Sebag since 2002.

9.1.5. Evaluation committees and invited expertise

- Antoine Cornuéjols, reviewer of the National research program ACI Masses de données.
- Cécile Germain-Renaud, reviewer of the National research program ACI Masses de données.
- Marc Schoenauer, reviewer for both ANR programs Young Researchers and Open Call ("appel blanc"); reviewer for the Austrian Science Fund; Professor evaluation for Profs P. Ross (Napier University), Fogarty (Galway University) and Riccardo Poli (IDSIA, Lugano, Switzerland).
- Michèle Sebag, reviewer for both ANR programs Young Researchers and Open Call ("appel blanc"); reviewer for RNTL; reviewer for the FNRS (Belgique); reviewer for the LINA Lab Nantes (CNRS).

9.1.6. Other evaluation activities

- Reviewer for PhD dissertation: Marc Schoenauer (2) ; Michèle Sebag (1) ; Antoine Cornuéjols (1).
- Reviewer for Habilitation: Michèle Sebag (1)

9.1.7. Summer schools, tutorials, invited seminars

- Michèle Sebag co-organized of the 50th Birthday of Artificial Intelligence, Nov. 3, 2006. Ministère de la Recherche (180 participants)
- Conference on AI (France Culture, Oct. 6, 2006; with Jean-Gabriel Ganascia, Luc Steels, Pierre-Yves Oudeyer), Michèle Sebag.

9.2. Enseignement

9.2.1. Defended doctorates

- Carlos Kavka, 6/7/06, Université Paris-Sud
- Vijay Pratap Singh, 18/12/06, Ecole des Mines

9.2.2. Graduate courses

- Master 2 Recherche (U. Paris-Sud), Data mining and machine learning (24 h): Michèle Sebag, Antoine Cornuéjols.
- Master 2 Recherche (U.Paris-Sud), Artificial and Natural Perception : Nicolas Bredeche (3h).
- Master 2 Recherche (U.Paris-Sud), Multi-agent Systems : Nicolas Bredeche (3h).
- Master 2 Recherche (U.Paris-Sud), Artificial Evolution and Robotics, Anne Auger, Nicolas Bredèche and Marc Schoenauer.
- Master 1 Recherche (ENS Cachan), Introduction to Machine Learning, Michèle Sebag (3h).

9.2.3. Other research-related teaching activities

- *Ecole Polytechnique*, Projects in Evolutionary Robotics in the *Modex d'Electronique*: Marc Schoenauer, Cédric Hartland.
- *Ecole Polytechnique*, Majeure "SEISM" (Engineering Science): one lesson (+ hands-on experiments) on Evolutionary Topological Optimum Design; chapter published in G. Allaire's book [12].
- *Ecole Polytechnique*, *Stages d'option*: Michèle Sebag, Marc Schoenauer.
- ENSTA (*Ecole Nationale Supérieure de Techniques Avancées*), in charge of the *Machine Learning* course: Antoine Cornuéjols.

10. Bibliography

Year Publications

Doctoral dissertations and Habilitation theses

[1] C. KAVKA. *Evolutionary Design of Geometric-Based Fuzzy Systems*, Ph. D. Thesis, Université Paris-Sud, 2006, http://tel.archives-ouvertes.fr/tel-00118883/en/.

[2] V. P. SINGH. *Automatic Seismic Velocity Inversion using Multi-Objective Evolutionary Algorithms*, Ph. D. Thesis, École des Mines de Paris, 2006, http://tel.archives-ouvertes.fr/tel-00120310/en/.

Articles in refereed journals and book chapters

[3] A. AUGER, P. CHATELAIN, K. P.. *R-leaping: Accelerating the stochastic simulation algorithm by reaction leaps*, in "J. Chem. Phys.", vol. 125, 2006.

[4] N. BREDECHE, Z. SHI, J.-D. ZUCKER. *Perceptual Learning and Abstraction in Machine Learning : an application to autonomous robotics*, in "IEEE Transactions on Systems, Man and Cybernetics, part C", vol. 36, n° 2, 2006, p. 172-181, http://hal.inria.fr/inria-00116923/en/.

[5] A. CORNUÉJOLS. *In order to learn: How the sequences of topics affect learning*, F. RITTER, J. NERB, E. LEHTINEN, T. O'SHEA (editors). , chap. Machine Learning: The Necessity of Order (is order in order?), Oxford University Press, 2006, http://hal.inria.fr/inria-00119757/en/.

[6] S. GELLY, O. TEYTAUD, N. BREDECHE, M. SCHOENAUER. *Universal Consistency and Bloat in GP*, in "Revue d'Intelligence Artificielle", vol. 20, n° 6, 2006, p. 805-827, http://hal.inria.fr/inria-00112840/en/.

[7] S. GELLY, O. TEYTAUD. *Bayesian Networks: a Non-Frequentist Approach for Parametrization, and a more Accurate Structural Complexity Measure*, in "Revue d'Intelligence Artificielle", vol. 20, n° 6, 2006, p. 717-755, http://hal.inria.fr/inria-00112838/en/.

[8] C. GERMAIN-RENAUD, C. LOOMIS, J. MOSCICKI, R. TEXIER. *Scheduling for Responsive Grids*, in "Journal of Grid Computing", 2006, http://hal.inria.fr/inria-00117486/en/.

[9] Y. GUERMEUR, O. TEYTAUD. *Younes Bennani, ed., Apprentissage Connexioniste*, chap. Estimation et contrôle des performances en généralisation des réseaux de neurones, Hermès, 2006.

[10] R. LERICHE, M. SCHOENAUER, M. SEBAG. *Modélisation Numérique: défis et perspectives, , Traité Mécanique et Ingénierie des Matériaux*, P. BREITKOPF, C. KNOPF-LENOIR (editors). , chap. Un état des lieux de l'optimisation évolutionnaire et de ses implications en sciences pour l'ingénieur, Hermès, 2006, http://hal.inria.fr/inria-00120733/en/.

[11] A. MOREAU, O. TEYTAUD, J.-P. BERTOGLIO. *Optimal estimation for Large-Eddy Simulation of turbulence and application to the analysis of subgrid models*, in "Physics of fluids", vol. 18, 2006.

[12] M. SCHOENAUER. *Optimisation évolutionnaire, in* G. ALLAIRE *: Conception optimale de structures*, Mathématiques et Applications, chap. , n° 58, Springer Verlag, 2006, p. 221-264.

[13] M. SEBAG. *Paradigmes et enjeux de l'informatique*, N. BIDOIT, L. F. DEL CERRO, S. FDIDA, B. VALLÉE (editors). , chap. Fouille de donnée, Hermès, 2006, p. 137-156.

Publications in Conferences and Workshops

[14] N. BASKIOTIS, M. SEBAG, M.-C. GAUDEL, S.-D. GOURAUD. *A Machine Learning approach for Statistical Software Testing*, in "Twentieth International Joint Conference on Artificial Intelligence, Hyderabad, India", 2006, http://hal.inria.fr/inria-00112681/en/.

[15] N. BASKIOTIS, M. SEBAG, M.-C. GAUDEL, S.-D. GOURAUD. *EXIST: Exploitation/Exploration Inference for Statistical Software Testing*, in "On-line Trading of Exploration and Exploitation, NIPS 2006 Workshop, Whistler, BC, Canada", 2006, http://hal.inria.fr/inria-00117172/en/.

[16] A. CORNUÉJOLS, F. THOLLARD. *Artificial data and language theory*, in "GI workshop 2006. Grammatical inference: workshop on open problems and new directions, 21/11/2005, Saint-Etienne, France", Colin de la Higuera, 2006, http://hal.inria.fr/inria-00119756/en/.

[17] A. DEVERT, N. BRÉDECHE, M. SCHOENAUER. *BlindBuilder : a new encoding to evolve Lego-like structures*, in "EUROGP 2006, Budapest, Hungary", Lecture Notes in Computer Science, vol. 3905, 2006, p. 61–72, http://hal.inria.fr/inria-00000995/en/.

[18] A. DEVERT, N. BREDECHE, M. SCHOENAUER. *Evolutionary Design of Buildable Objects with BlindBuilder : an Empirical Study*, in "Asia-Pacific Workshop on Genetic Programming, Hanoi, Vietnam", Proceedings of the Third Asian-Pacific workshop on Genetic Programming, The Long Pham and Hai Khoi Le and Xuan Hoai Nguyen, 2006, p. 98–109, http://hal.inria.fr/inria-00118652/en/.

[19] C. GAGNÉ, M. SCHOENAUER, M. PARIZEAU, M. TOMASSINI. *Genetic Programming, Validation Sets, and Parsimony Pressure*, in "EuroGP 2006, Budapest, Hongrie", P. C. ET AL. (editor). , Lecture Notes in Computer Science, vol. 3905, Springer Verlag, 2006, p. 109-120, http://hal.inria.fr/inria-00000996/en/.

[20] C. GAGNÉ, M. SCHOENAUER, M. SEBAG, M. TOMASSINI. *Genetic Programming for Kernel-based Learning with Co-evolving Subsets Selection*, in "Parallel Problem Solving from Nature, Reykjavik", T. R. ET AL. (editor). , LNCS, n^{o} 4193, Springer Verlag, 2006, p. 1008-1017, http://hal.inria.fr/inria-00116344/en/.

[21] S. GELLY, J. MARY, O. TEYTAUD. *Learning for stochastic dynamic programming*, in "11th European Symposium on Artificial Neural Networks (ESANN), bruges Belgium", 2006, http://hal.inria.fr/inria-00112796/en/.

[22] S. GELLY, J. MARY, O. TEYTAUD. *On the ultimate convergence rates for isotropic algorithms and the best choices among various forms of isotropy*, in "Parallel Problem Solving from Nature, Reykjavik", LNCS, n^{o} 4193, 2006, p. 32-41, http://hal.inria.fr/inria-00112816/en/.

[23] S. GELLY, S. RUETTE, O. TEYTAUD. *Comparison-based algorithms: worst-case optimality, optimality w.r.t a bayesian prior, the intraclass-variance minimization in EDA, and implementations with billiards*, in "Parallel Problem Solving from Nature BTP-Workshop, Reykjavik", 2006, http://hal.inria.fr/inria-00112813/en/.

[24] S. GELLY, O. TEYTAUD, C. CAGNE. *Resource-Aware Parameterizations of EDA*, in "Congress on Evolutionary Computation, Vancouver, BC, Canada", 2006, http://hal.inria.fr/inria-00112803/en/.

[25] S. GELLY, O. TEYTAUD. *OpenDP a free Reinforcement Learning toolbox for discrete time control problems*, in "NIPS Workshop on Machine Learning Open Source Software, Whistler (B.C.)", 2006, http://hal.inria.fr/inria-00117392/en/.

[26] C. GERMAIN-RENAUD. *Scheduling for Interactive Grids*, in "First EGEE User Forum, Genève/Suisse", 2006, http://hal.inria.fr/inria-00117492/en/.

[27] C. GERMAIN-RENAUD, C. LOOMIS, R. TEXIER, A. OSORIO. *Grid Scheduling for Interactive Analysis*, in "HealthGrid 2006, Valencia/Spain", in: Studies in Health Technology and Informatics, Challenges and Opportunities of Health Grids, vol. 120, IOS Press, 2006, p. 25-33, http://hal.inria.fr/inria-00117491/en/.

[28] C. HARTLAND, N. BREDECHE. *Evolutionary Robotics: From Simulation to the Real World using Anticipation*, in "ABIALS, Rome/Italie", Oui, 2006, http://hal.inria.fr/inria-00120115/en/.

[29] C. HARTLAND, N. BREDÈCHE. *Evolutionary Robotics, Anticipation and the Reality Gap*, in "ROBIO, Kunming/Chine", Oui, 2006, http://hal.inria.fr/inria-00120116/en/.

[30] T. HEITZ, J. AZÉ, M. ROCHE, A. MELA, P. PEINL, M. AMAR DJALIL. *Présentation de DEFT 06 (DÉfi Fouille de Textes)*, in "Atelier DEFT'06 - SDN'06 (Semaine du Document Numérique), Fribourg, Suisse", Actes de l'atelier DEFT'06, SDN'06 (Semaine du Document Numérique), 2006, p. 1-10, http://hal.inria.fr/inria-00119612/en/.

[31] T. HEITZ. *Modélisation du prétraitement des textes*, in "JADT'06 (International Conference on Statistical Analysis of Textual Data), Besançon, France", Proceedings of JADT'06, vol. 1, 2006, p. 499-506, http://hal.inria.fr/inria-00119608/en/.

[32] L. HUGUES, N. BREDECHE. *Simbad : an Autonomous Robot Simulation Package for Education and Research*, in "Simulation of Adaptive Behavior (SAB 2006), Rome, Italy", 2006, http://hal.inria.fr/inria-00116929/en/.

[33] V. KRMICEK, M. SEBAG. *Functional Brain Imaging with Multi-Objective Multi-Modal Evolutionary Optimization*, in "Parallel Problem Solving from Nature, Reykjavik", T. R. ET AL. (editor). , LNCS, n^{O} 4193, Springer Verlag, 2006, p. 382-391, http://hal.inria.fr/inria-00116342/en/.

[34] S. LALLICH, O. TEYTAUD, E. PRUDHOMME. *Association rules interestingness: measure and validation*, in "Quality Measures in Data Mining", F. GUILLET, H.-J. HAMILTON (editors). , Springer, 2006, 23.

[35] S. LALLICH, O. TEYTAUD, E. PRUDHOMME. *Statistical inference and data mining: false discoveries control*, in "proceedings of the 17th COMPSTAT Symposium of the IASC", 2006.

[36] M. SCHOENAUER, P. SAVÉANT, V. VIDAL. *Divide-and-Evolve : une nouvelle méta-heuristique pour la planification temporelle indépendante du domaine*, in "Journées Francophones Planification, Décision, Apprentissage, Toulouse", F. GARCIA, G. VERFAILLIE (editors). , GDR I3 groupe PDMIA, 2006, http://hal.inria.fr/inria-00121779/en/.

[37] M. SCHOENAUER, P. SAVÉANT, V. VIDAL. *Divide-and-Evolve: a New Memetic Scheme for Domain-Independent Temporal Planning*, in "EvoCOP2006, Budapest", J. GOTTLIEB, G. RAIDL (editors). , LNCS, n^{O} 3906, Springer Verlag, 2006, p. 247-260, http://hal.inria.fr/inria-00000975/en/.

[38] Y. SEMET, M. SCHOENAUER. *On the Benefits of Inoculation, an Example in Train Scheduling*, in "GECCO-2006, Seattle", M. C. ET AL. (editor). , ACM Press, 2006, http://hal.inria.fr/inria-00116345/en/.

[39] O. TEYTAUD, S. GELLY. *General lower bounds for evolutionary algorithms*, in "Parallel Problem Solving from Nature, Reykjavik", LNCS, n^{O} 4193, 2006, p. 21-31, http://hal.inria.fr/inria-00112820/en/.

[40] O. TEYTAUD. *How entropy-theorems can show that approximating high-dim Pareto-fronts is too hard*, in "Bridging the Gap between Theory and Practice - Workshop PPSN-BTP", 2006.

[41] O. TEYTAUD. *Why Simulation-Based Approachs with Combined Fitness are a Good Approach for Mining Spaces of Turing-equivalent Functions*, in "Proc. of the IEEE Congress on Evolutionary Computation (CEC 2006)", 2006.

Internal Reports

[42] S. GELLY, Y. WANG, R. MUNOS, O. TEYTAUD. *Modification of UCT with Patterns in Monte-Carlo Go*, Rapport de recherche INRIA, n^{O} RR-6062, 2006, http://hal.inria.fr/inria-00117266/en/.

[43] D. TESSIER, M. SCHOENAUER, C. BIERNACKI, G. CELEUX, G. GOVAERT. *Evolutionary Latent Class Clustering of Qualitative Data*, Rapport de recherche INRIA, n^{o} RR-6082, 2006, http://hal.inria.fr/inria-00122088/en/.

Miscellaneous

[44] M. AMIL, C. GAGNÉ, N. BREDÈCHE, S. GELLY, M. SCHOENAUER, O. TEYTAUD. *How to ensure universal consistency and no bloat with VC-dimensionDagstuhl Seminar "Theory of Evolutionary Algorithms", 06061*, 2006.

[45] S. GELLY, Y. WANG. *Exploration-Exploitation in Go: UCT for Monte-Carlo-Go. On-line trading of Exploration and Exploitation Workshop, NIPS Conference*, 2006.

[46] C. HARTLAND, S. GELLY, N. BASKIOTIS, O. TEYTAUD, M. SEBAG. *Multi-armed Bandit, Dynamic Environments and Meta-BanditsOnline Trading of Exploration and Exploitation Workshop, NIPS*, 2006, http://hal.archives-ouvertes.fr/hal-00113668/en/.

[47] O. TEYTAUD, S. GELLY, S. LALLICH, E. PRUDHOMME. *Quasi-random resamplings, with applications to rule extraction, cross-validation and (su-)bagging. Pascal Workshop IIIA'2006*, 2006.