Master Thesis Proposal

Temporal Data Mining: Learning from Positive Examples

Oded Maler, Nicolas Basset, VERIMAG, Grenbole @univ-grenoble-alpes.fr

The ultimate goal of the project is to build small models that capture the dynamics of complex and high-dimensional systems, based on sample observations of their behaviors. For example, the system can be an analog circuit C with thousands of transistors and the small abstract model C' will be a low-dimensional differential equation that generates observable behaviors close to those of C. The reduced model can be exported as an approximate description of C and be used for low-cost simulation and verification with other components of the whole system. Another example, coming from our collaboration with Toyota, is a complex car engine which is modeled as a huge network of Matlab/Simulink components. By simulating the system with various input scenarios (driver behavior, road conditions) we obtain a sample S of behaviors (trajectories, signals) for which we would like to build a succinct description, expressed for example as a formula in *signal temporal logic* (STL) or a timed regular expression.

In the context of machine learning, this is a problem of learning from *positive examples*, because we do not observe impossible behaviors. If all we want from C' is to be compatible with S, we have two extreme choices: either let C' be the system that generates/accepts exactly S (known as over-fitting in machine learning) or let C' be the universal system that accepts everything. Both choices are, of course, unsatisfactory and we should choose some other function/set in the sub-lattice between the two.

The project will study this issue first on Boolean functions, where the problem resembles some variant of logic minimization and is concerned with fundamental computer science objects such as terms/cubes, DNF formulas and their sampling. The idea is to develop a learning algorithm that explores the trade-offs between the cardinality of C' and its descriptive complexity (the number of terms in its DNF representation).

After solving the problem for static functions, we will move to functions over Boolean sequences, with the goal of producing a sample-compatible formula in linear-time temporal logic (LTL) or a regular expression. A finite sequence of length k over $\mathbb{B}^n$ can be represented using nk Boolean variables of the form $x_i[t]$. Hence a set/language of sequences can be written as a Boolean function over these variables. One possible research direction would be to characterize temporal functions (sequence classification), those that are expressed compactly by sequential formalisms, as special classes of Boolean functions closed under some variable permutations and not others. Similar investigations can be applied to spatial functions (image classification). The mix of theory and practice (implementation, empirical evaluation) will be determined by the inclination of the student.

Further Reading:

Shai Shalev-Shwartz and Shai Ben-David, Understanding Machine Learning: From Theory to Algorithms,

Do not hesitate to contact us for clarifications.

The work will be conducted at VERIMAG, an internationally recognized lab in verification, embedded systems and hybrid (discrete-continuous, cyber-physical) systems. VERIMAG is located in the nice campus of UGA (Universite Grenobles-Alpes. There will be a possibility to continue for a thesis.