PAC Learning of Thomas Regulatory Networks from Time-Series Data

Arthur Carcano[1], François Fages[2], Jérémy Grignard[2], and Sylvain Soliman[2]

[1] Ecole Normale Supérieure, Paris, France
arthur.carcano@ens.fr
[2] Inria, University Paris-Saclay, Lifeware group, France
Francois.Fages@inria.fr, Sylvain.Soliman@inria.fr

Abstract. Automating the process of model building from experimental data is a very desirable goal to palliate the lack of modellers for many applications. However, despite the spectacular progress of machine learning techniques in data analytics, classification, clustering and prediction making, learning dynamical models from data time-series is still challenging. In this paper we investigate the use of the Probably Approximately Correct (PAC) learning framework of Leslie Valiant as a method for the automated discovery of influence models of biochemical processes from Boolean and stochastic traces. We show that Thomas' Boolean influence systems can be naturally represented by k-CNF formulae and learned from time-series data with a quasi linear number of Boolean activation samples per species, and that positive Boolean influence systems can be represented by monotone DNF formulae and learned actively with both activation samples and oracle calls. We evaluate the performance of this approach on a model of T-lymphocyte differentiation, with and without prior knowledge, and discuss its merits as well as its limitations with respect to realistic experiments.

1 Introduction

Modelling biological systems is still an art which is currently limited in its applications by the number of available modellers. Automating the process of model building is thus a very desirable goal to attack new applications, develop patient-tailored therapeutics, and also design experiments that can now be largely automated with a gain in both the quantification and the reliability of the observations, at both the single cell and population levels.

Machine learning is revolutionising the statistical methods in biological data analytics, data classification and clustering, and prediction making. However, learning dynamical models from data time-series is still challenging. A recent survey on probabilistic programming [9] highlighted the difficulties associated with modelling time, and concluded that existing frameworks are not sufficient in their treatment of dynamical systems. There has been early work on the use of machine learning techniques, such as inductive logic programming [12] combined with active learning in the vision of the "robot scientist" [4], to infer gene

functions, metabolic pathway descriptions [1,2] or gene influence systems [3], or to revise a reaction model with respect to CTL properties [5]. Since a few years, progress in this field is measured on public benchmarks of the "Dream Challenge" competition [11]. Logic Programming, and especially *Answer Set Programming* (ASP), provide efficient tools such as CLASP [7] to implement learning algorithms for Boolean models. They have been applied in [8] to the detection of inconsistencies in large biological networks, and have been subsequentially applied to the inference of gene networks from gene expression data and to the design of discriminant experiments [19]. Furthermore, ASP has been combined with CTL model-checking in [13] to learn mammalian signalling networks from time series data, and identify erroneous time-points in the data.

In this paper, we consider the framework of Probably Approximately Correct (PAC) Learning which was introduced by Leslie Valiant in his seminal paper on a theory of the learnable [17]. Valiant questioned what can be learned from a computational viewpoint, and introduced the concept of PAC learning, together with a general-purpose polynomial-time learning protocol. Beyond the algorithms that one can derive with this methodology, Valiant's theory of the learnable has profound implications on the nature of biological and cognitive processes, of collective and individual behaviors, and on the study of their evolution [18].

Here we investigate PAC learning as a method for the automated discovery of influence models of biochemical processes from time-series data. To the best of our knowledge, the application of PAC learning to dynamical models of biochemical systems has not been reported before. We show that Thomas' gene regulatory networks [16,15] can be naturally represented by Boolean formulae in conjunctive normal forms with a bounded number of litterals (i.e. k-CNF formulae), and can be learned from Boolean transitions with a quasi linear number of Boolean transition samples, using Valiant's PAC learning algorithm for k-CNF formulae. We also show that Boolean influence systems with their positive Boolean semantics discussed in [6] can be naturally represented by monotone DNF formulae, and learned actively from a set of positive samples with calls to an oracle. These results[3] are evaluated on a Boolean influence model of the differentiation of the T-helper lymphocytes from [14,10], composed of 32 influences and 12 variables.

2 Valiant's PAC Learning Algorithms

Let n be the dimension of the model to learn, and let us consider a finite set of Boolean variables $x_1, \ldots, x_n$, A vector is an assignment of the n variables to $\mathbb{B}_* = \{0, 1, *\}$; A total vector is a Boolean assignment, in $\mathbb{B} = \{0, 1\}$; A Boolean function $G : \mathbb{B}^n \to \mathbb{B}$; assigns a Boolean value to each total vector; A concept $F : {\mathbb{B}_*}^n \to \mathbb{B}$ assigns a Boolean value to each vector.

[3] The code is available at `http://lifeware.inria.fr/wiki/software/#CMSB17`.

The idea behind the PAC learning protocol is to discover a concept, or a Boolean function, G which approximates a hidden concept F, while restricting oneself to the two following operations :

- SAMPLE(): returns a positive example, i.e. a vector v such that $F(v) = 1$. The output of SAMPLE() is assumed to follow a given probability distribution $D(v)$, which is used to measure the approximation of the result.
- ORACLE(v): returns the value of $F(v)$ for any input vector v.

Definition 1 ([17]). *A class $\mathcal{M}$ of* Boolean functions *is said to be* learnable *if there exists an algorithm $\mathcal{A}$ with some precision parameter $h \in \mathbb{N}$ such that $\mathcal{A}$ runs in polynomial time both in n and h; and for any function F in $\mathcal{M}$, and any distribution D on the positive examples, $\mathcal{A}$ deduces with probability higher than $1 - h^{-1}$ an approximation G of F such that*

- *$G(v) = 1$ implies $F(v) = 1$ (no false positive)*
- $\sum_{v \text{ s.t. } F(v)=1 \wedge G(v)=0} D(v) < h^{-1}$ *(low probability of false negatives)*

Valiant showed the learnability of some important classes of functions in this framework, in particular for Boolean formulae in conjunctive normal forms with at most k literals per conjunct (k-CNF), and for monotone (i.e. negation free) Boolean formulae in disjunctive normal form (DNF). The computational complexities of the PAC learning algorithms are expressed in terms of some function

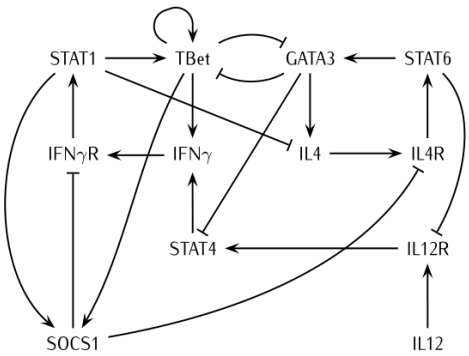
Theorem 1 ([17]). *For any k, the class of k-CNF formulae (i.e. with at most k literals per conjunct) on n variables is learnable with an algorithm that uses $L(h, (2n)^{k+1})$ positive examples and no call to the oracle (where $L(h, S) \leq 2h(S + \log_e h)$). The class of monotone (i.e. without negation) DNF formulae on n variables is also learnable with an algorithm that uses $L(h, d)$ examples and dn calls to the oracle (where d is the largest number of prime implicants in an equivalent rewriting of the formula to learn as a non-redundant sum of prime-implicants).*

3 Thomas's Boolean Regulatory Network

Definition 2 ([15]). *A* Thomas *network on a finite set of genes $\{x_1, \ldots, x_n\}$ is defined by n Boolean functions $\{f_1, \ldots, f_n\}$ which give for each gene its possible next state, given the current state.*

k-CNF formulae can be used to represent Thomas gene regulatory network functions with some reasonable restrictions on their connectivity, e.g. for networks of degree bounded by k, When restricting to monotone activation functions, i.e. without negation testing the absence of expression of a gene, monotone DNF formulae can be used as well to represent monotone Thomas networks, i.e. with positive and negative influences but no influence inhibitors [6].

Here we consider a regulatory network of 12 variables which models the differentiation between Th-1 and Th-2 lymphocytes from an original CD4+ T helper (Th-0). The model has three different stable states corresponding to Th-0 (naive lymphocyte), Th-1 and Th-2 when IL12 is off, and two others when IL12 is on (the Th-0 one is lost). This model, presented in [14] is actually a Boolean simplification of the original multi-level model of [10].



4 PAC Learning from Boolean and Stochastic Traces

Valiant's work on PAC learning provides an elegant trail to attack the challenge of inferring the structure of influence models from the observation of data time series, and more precisely to automatically discover possible regulatory networks of a biochemical process, given sufficiently precise observations of its executions.

The Boolean dynamics of biochemical influence systems, including Thomas regulatory networks, can be represented by k-CNF formulae without loss of generality, and k-CNF PAC learning algorithm can be used to infer the strucutre of the network from a sufficiently large and diverse set of state transition traces. When dimension increases, we show on the example of T-lymphocyte differentiation from the litterature that the k-CNF PAC learning algorithm can also leverage available prior knowledge on the system to deliver precise results with a reasonable amount of data.

The Boolean dynamics of positive influence systems can also be straightforwardly represented by monotone DNF activation and deactivation functions, and monotone DNF PAC learning algorithm applied with an interesting recourse to oracles which are particularly relevant in the perspective of online active learning and experimental design.

More work is needed however to make comparisons on common benchmarks with other approaches already investigated in this context, such as Answer Set Programming (ASP) and budgeted learning, and to investigate the applicability of these methods to real experiments taking into account particular biological technologies.

References

1. Angelopoulos, N., Muggleton, S.H.: Machine learning metabolic pathway descriptions using a probabilistic relational representation. Electronic Transactions in Artificial Intelligence 7(9) (2002), also in Proceedings of Machine Intelligence 19
2. Angelopoulos, N., Muggleton, S.H.: Slps for probabilistic pathways: Modeling and parameter estimation. Tech. Rep. TR 2002/12, Department of Computing, Imperial College, London, UK (2002)

3. Bernot, G., Comet, J.P., Richard, A., Guespin, J.: A fruitful application of formal methods to biological regulatory networks: Extending Thomas' asynchronous logical approach with temporal logic. Journal of Theoretical Biology 229(3), 339–347 (2004)
4. Bryant, C.H., Muggleton, S.H., Oliver, S.G., Kell, D.B., Reiser, P.G.K., King, R.D.: Combining inductive logic programming, active learning and robotics to discover the function of genes. Electronic Transactions in Artificial Intelligence 6(12) (2001)
5. Calzone, L., Chabrier-Rivier, N., Fages, F., Soliman, S.: Machine learning biochemical networks from temporal logic properties. In: Plotkin, G. (ed.) Transactions on Computational Systems Biology VI, Lecture Notes in BioInformatics, vol. 4220, pp. 68–94. Springer-Verlag (Nov 2006)
6. Fages, F., Martinez, T., Rosenblueth, D., Soliman, S.: Influence systems vs reaction systems. In: E. Bartocci, P. Lio, N.P. (ed.) CMSB'16: Proceedings of the fourteenth international conference on Computational Methods in Systems Biology. Lecture Notes in BioInformatics, vol. 9859, pp. 98–115. Springer-Verlag (Sep 2016)
7. Gebser, M., Kaufmann, B., Neumann, A., Schaub, T.: clasp: A conflict-driven answer set solver. In: In Proc. LPNMR'07. pp. 260–265. Springer (2007)
8. Gebser, M., Schaub, T., Thiele, S., Usadel, B., Veber, P.: Detecting inconsistencies in large biological networks with answer set programming. In: de la Banda, M.G., Pontelli, E. (eds.) ICLP'08, Proceedings of the 24th International Conference on Logic Programming. Lecture Notes in Computer Science, vol. 5366, pp. 130–144. Springer-Verlag (2008)
9. Gordon, A.D., Henzinger, T.A., Nori, A.V., Rajamani, S.K.: Probabilistic programming. In: Proceedings of the on Future of Software Engineering. pp. 167–181. FOSE 2014, ACM, New York, NY, USA (2014)
10. Mendoza, L.: A network model for the control of the differentiation process in Th cells. Biosystems 84(2), 101–114 (2006)
11. Meyer, P., Cokelaer, T., Chandran, D., Kim, K.H., Loh, P.R., Tucker, G., Lipson, M., Berger, B., Kreutz, C., Raue, A., Steiert, B., Timmer, J., Bilal, E., Sauro, H.M., Stolovitzky, G., Saez-Rodriguez, J.: Network topology and parameter estimation: from experimental design methods to gene regulatory network kinetics using a community based approach. BMC Systems Biology 8(1), 1–18 (2014)
12. Muggleton, S.H.: Inverse entailment and progol. New Generation Computing 13, 245–286 (1995)
13. Ostrowski, M., Paulevé, L., Schaub, T., Siegel, A., Guziolowski, C.: Boolean network identification from perturbation time series data combining dynamics abstraction and logic programming. Biosystems 149, 139–153 (2016)
14. Remy, E., Ruet, P., Mendoza, L., Thieffry, D., Chaouiya, C.: From Logical Regulatory Graphs to Standard Petri Nets: Dynamical Roles and Functionality of Feedback Circuits, pp. 56–72. Springer-Verlag, Berlin, Heidelberg (2006)
15. Thomas, R.: Boolean formalisation of genetic control circuits. Journal of Theoretical Biology 42, 565–583 (1973)
16. Thomas, R.: Regulatory networks seen as asynchronous automata : a logical description. Journal of Theoretical Biology 153, 1–23 (1991)
17. Valiant, L.: A theory of the learnable. Communications of the ACM 27(11), 1134–1142 (1984)
18. Valiant, L.: Probably Approximately Correct. Basic Books (2013)
19. Videla, S., Konokotina, I., Alexopoulos, L.G., Saez-Rodriguez, J., Schaub, T., Siegel, A., Guziolowski, C.: Designing experiments to discriminate families of logic models. Frontiers in Bioengineering and Biotechnology 3, 131 (2015)