Automatic Identification of Hierarchy in Multivariate Data

Ilknur Icke
Morphology, Evolution and Cognition Lab
Department of Computer Science
University of Vermont
Burlington, VT 05401
ilknur.icke@uvm.edu

Joshua C. Bongard
Morphology, Evolution and Cognition Lab
Department of Computer Science
University of Vermont
Burlington, VT 05401
jbongard@uvm.edu

ABSTRACT

Given n variables to model, symbolic regression (SR) returns a flat list of n equations. As the number of state variables to be modeled scales, it becomes increasingly difficult to interpret such a list. Here we present a symbolic regression method that detects and captures hidden hierarchy in a given system. The method returns the equations in a hierarchical dependency graph, which increases the interpretability of the results. We demonstrate two variations of this hierarchical modeling approach, and show that both consistently outperform non-hierarchical symbolic regression on a number of synthetic data sets.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: [Database applications-Data Mining]

General Terms

Algorithms

Keywords

system identification, hierarchy, symbolic regression

1. SYMBOLIC REGRESSION APPROACH TO MODEL HIERARCHY

Hierarchy is thought to be a fundamental characteristic of many complex systems such as biological organisms [4], ecological systems [2], the internet, traffic networks [3] and, arguably, social organizations [1]. In this paper, we propose symbolic regression based approaches to model hierarchical relationships in multivariate data. A simple approach is to model each variable separately in terms of all other variables using SR (the naive symbolic regression approach, figure 1). The best models are identified based on the error on validation data, then, the variables that appear in these models are identified as the predictors for each respective modeled variable. An adjacency matrix is then built based on these identified predicted variable-predictor mappings. Finally, the algorithm returns the adjacency matrix representing the connectivity between the variables along with the set of best models evolved for each non-stimulus variable. By modeling each variable independently, this algorithm does not impose any constraints on the connectivity.

GECCO'13 Companion, July 6–10, 2013, Amsterdam, The Netherlands.
ACM 978-1-4503-1964-5/13/07.

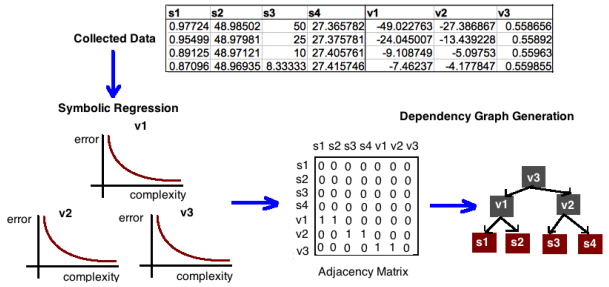

Figure 1: The naive symbolic regression (NSR) algorithm. Leaves: stimuli (independent variables) and internal nodes: state variables (dependent variables)

We also devised an algorithm that enforces hierarchical extraction of the dependencies in an iterative manner. At each iteration, dependencies for one non-stimulus variable are discovered using SR. The algorithm starts with only the stimuli as the set of available independent variables. After first iteration, the variable (v_i) that is best explained by a subset of these inputs is determined. Then, all predictors for variable (v_i) are removed from the set of available independent variables in accordance with our constraint that inputs can not overlap. Next, v_i is added to the list of independent variables. The algorithm stops after each (non-stimulus) variable has been modeled using SR. Figure 2 shows two variants of this hierarchical approach in which the selection of the best modeled variable and the corresponding predictors are performed in two different ways.

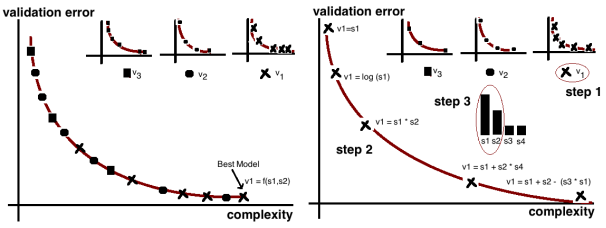

Figure 2: (left panel) The HSR version 1 (h1). Best model is chosen on the combined Pareto set, from which the predictors are extracted. (right panel) The HSR version 2 (h2). The state variable with the best error-complexity trade-off is selected, the most frequent predictors are extracted. An extra SR step with the chosen predictors generates the model for that state variable

2. EXPERIMENTAL RESULTS

All three algorithms were compared on a synthetic benchmark data suite, where a number of hidden target hierarchical systems with varying arities and tree heights were randomly generated. Each expression contained only {+,-,* and protected / } operators without any constant values. The degree of nonlinearity was kept constant by enforcing that only binary nonlinear interactions are allowed in the target expressions. For instance, the hidden expression for a state variable can be $v_1 = s_1 + s_2 - s_4 * s_5 - s_3$, but not $v_1 = s_1 * s_2 / s_4 * s_5 - s_3$. All datasets were divided into training, validation and testing partitions.

Figure 3 shows the results for a number of arity/tree height configurations where 30 hidden target systems were randomly generated per configuration. Each algorithm was given a 10-minute run-time budget per dataset. For the percentage of correct edges, each pair of algorithms were compared using the left-tailed Wilcoxon rank sum test with Bonferroni correction and unequal variances assumption. The cases where the h1 and h2 algorithms are significantly better than the naive algorithm, and when the h2 algorithm is significantly better than the h1 algorithm, are presented using $*$ sign ($***$:$\alpha = 0.001$, $**$:$\alpha = 0.01$, $*$:$\alpha = 0.05$). For the test set error results, we performed right-tailed Wilcoxon rank sum tests, since lower test error indicates better performance. The heatmaps show error versus number of correctly identified edges where it is evident that the naive algorithm mostly finds low-error models at the expense of missing many edges. The problem becomes increasingly difficult for all three approaches as the arity increases and the hierarchical approach no longer outperforms the naive for arity=5.

3. CONCLUSION

Extracting and visualizing the relationships in a hierarchical system as a dependency graph improves the intelligibility of the overall model, compared to the flat list of equations produced by traditional symbolic regression. Our results clearly show that in order to find hierarchy, one needs to explicitly search for it rather than waiting for the hierarchical models to emerge in an unconstrained search such as in the naive SR. On the other hand, it is possible to devise many ways for explicitly searching for hierarchy in the data. In this paper, two such approaches were explored. The focus of our current work is to further explore more efficient ways for the selection process at each stage and to extend the algorithm to model more general systems that exhibit mixtures of hierarchy and network connectivity.

4. REFERENCES

[1] S.V Buldyrev, N.V Dokholyan, S Erramilli, M Hong, J.Y Kim, G Malescio, and H.E Stanley. Hierarchy in social organization. *Physica A: Statistical Mechanics and its Applications*, 330(3-4):653 – 659, 2003.

[2] Hironori Hirata and Robert E. Ulanowicz. Information theoretical analysis of the aggregation and hierarchical structure of ecological networks. *Journal of Theoretical Biology*, 116(3):321 – 341, 1985.

[3] Erzsébet Ravasz and Albert-László Barabási. Hierarchical organization in complex networks. *Phys. Rev. E*, 67:026112, Feb 2003.

[4] Ralf J Sommer. Homology and the hierarchy of biological systems. *Bioessays*, 30(7):653–8, 2008.

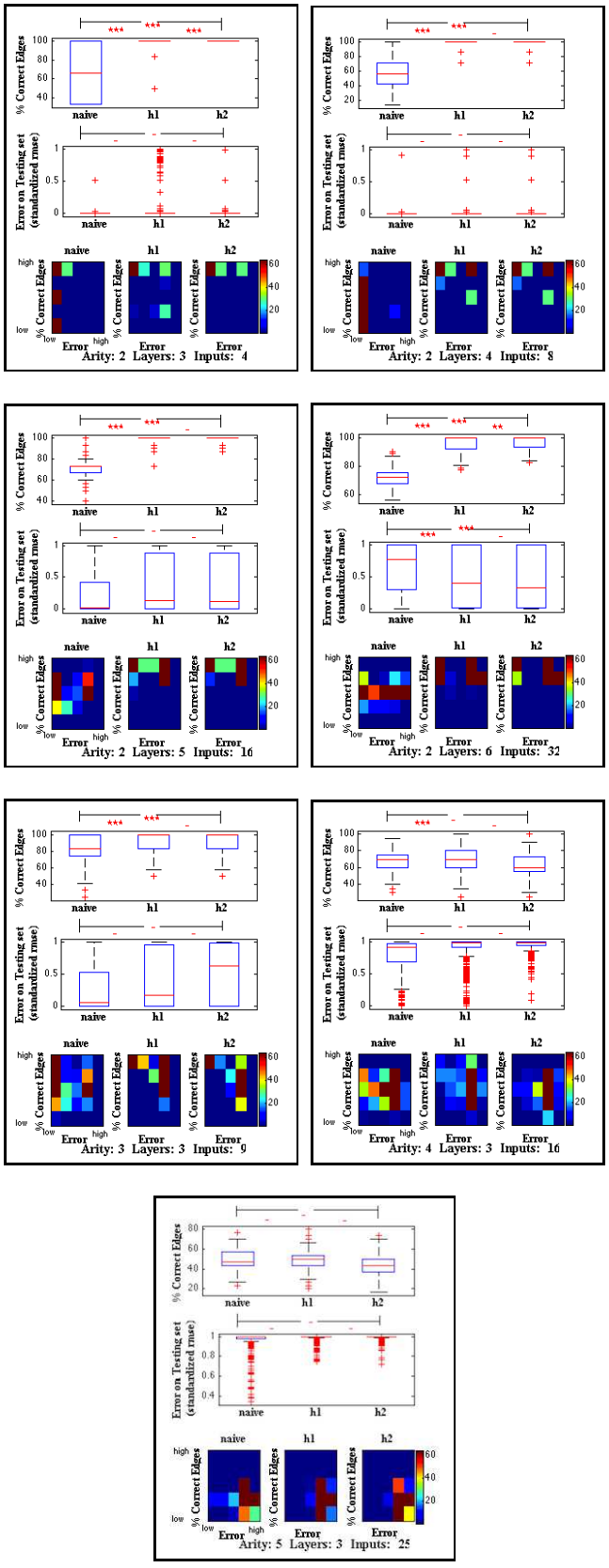

Figure 3: For binary systems (arity=2), both hierarchical approaches outperform the naive approach even when the tree height increases. The hierarchical approach outperforms the naive approach until arity=5