Model Validation: A statistical-based criterium of hypotheses acceptance in numerical reasoning

Alexessander Alves, Rui Camacho and Eugénio Oliveira

LIACC, Rua do Campo Alegre, 823, 4150 Porto, Portugal
FEUP, Rua Dr Roberto Frias, 4200-465 Porto, Portugal
alves@ieee.org {rcamacho,eco}@fe.up.pt
tel: +351 22 508 1400 fax: +351 22 508 1443

Abstract. Current ILP systems that perform numerical reasoning, select the best hypothesis using exclusively the scored value of the cost function. The cost function, by itself, cannot guarantee the *goodness-of-fit* of the induced hypotheses in numerical domains. Consequently the induced theory may not capture the overall structure of the underlying process that generated data. This paper proposes a statistical-based criterion for hypotheses acceptance, called model validation, that assess the *goodness-of-fit* of the induced hypotheses in numerical domains.
We have found this extension essential to improve on results over ML and statistical-based algorithms used in the empirical evaluation study.

1 Introduction

Current ILP approaches [1] to numerical domains usually carry out a search through the hypothesis space looking for a minimal value of a cost function like the Root Mean Square Error (RMSE). The RMSE is traditionally the hypothesis evaluation method and the minimum coverage of an hypothesis is the only acceptance criterium used on those systems. Systems like TILDE [2], LAGRAMGE [3], and FORS [4] are of that kind.

The minimisation of distance-based criteria like RMSE do not guarantee, by itself, the *goodness-of-fit* of the induced hypothesis. Which means that even hypotheses that score a small error on the training examples, may perform poorly on test data not only due to the over fitting problem, but also due to the fact that these hypotheses may not capture the overall structure of the underlying process that generated data. Thus, one problem with the minimisation of RMSE in noisy domains is that the models tend to be brittle, since the error is small when covering a small number of examples. The end result is a large set of clauses [1] to cover the complete set of examples, which is also a drawback on the intelligibility of ILP induced models.

A technique to mitigate this problem is the inclusion of statistical tests for model validation in the search process, namely as part of the hypothesis acceptance criterion of the ILP algorithm. Statistical tests may assess the correctness

[1] In this paper terms hypothesis and clause are used interchangeably

of the functional part of a model, assure that no significant terms will be missing and no systematic error or process drifting will exist. They may also check if the error structure follows an assumed distribution like for example the Gaussian distribution.

In this paper, we propose improvements on the numerical reasoning capabilities of ILP systems by introducing a new step on the basic ILP cycle algorithm that consists of an hypothesis acceptance criterium based on statistical tests for model validation.

We have made the empirical evaluation of the proposal in datasets collected from statistics and time series literature. For each dataset we compare our results with the published ones. The experimental results show that an ILP system extended with such procedures compares very well with statistical methods.

The rest of the paper is organised as follows. Section 2 identifies the steps of a basic ILP algorithm that are subject to the improvements proposed in this paper. The proposals for Model Validation are discussed in Section 3. Section 4 presents the experimental findings. Finally, in Section 5 we draw the conclusions.

2 Search Improvements

To establish a context for our proposals consider the main cycle of a general purpose ILP system as depicted in Algorithm 1. In ILP, the search procedure is usually an iterative greedy set-covering algorithm that finds the best clause on each iteration and removes the covered examples. Each hypothesis generated during the search is evaluated to determine their quality. A widely used approach in numerical domains is to score a hypothesis by measuring its Mean Absolute Error (MAE) or its RMSE. Algorithm 1 presents an overview of the procedure and identifies the step (step 4) modified by our proposal.

Algorithm 1 Basic cycle of a greedy set-covering ILP algorithm

```
1: repeat
2:     repeat
3:         Synthesise an hypothesis
4:         Accept a hypothesis (Model Validation)
5:         Evaluate hypothesis (Model Selection)
6:         Update best hypothesis
7:     until Stopping Criterion satisfied
8:     Remove explained examples
9: until "All" examples explained
```

Assume that the methodology for generating hypothesis considered in step 3 of algorithm 1 assures a complete search of the hypothesis space. We propose an improvement to step 4 where an hypothesis is checked if it is a satisfactory approximation of the underlying process that generated data, only hypotheses that are not rejected on that step are evaluated on the Model Selection step.

The hypothesis evaluation method used on the Model Selection step (5) is the RMSE, although any other method may be used for selecting the *best* hypothesis, including score functions that are hypothesis complexity aware. The stopping criterium is a predefined parameter that limits the maximum number of nodes explored.

3 Model Validation

In most applications, the true nature of the model is unknown. It is therefore of fundamental importance to assess the goodness-of-fit of each conjectured hypothesis. This is performed in a step of the induction process called *Model Validation*. Model Validation allows the system to check if the hypothesis is indeed a satisfactory model of the data. This step is common in both Machine Learning and Statistical Inference.

There are various ways of checking if a model is satisfactory. The most common approach is to examine the residuals. The residuals are the random process formed from the differences between the observed and predicted values of a variable. The behaviour of the residuals may be used to check the adequacy of the fitted model as a consequence of the Wold's theorem, defined as follows.

Theorem 1 (Wold's Theorem). *Any real-valued stationary process may be decomposed into two different parts. The first is totally deterministic. The second totally stochastic. The stochastic part of the process may be written as a sequence of serially uncorrelated random variables z with zero mean and variance σ^2. The stationarity condition imply $\sigma < \infty$, thus z is a White Noise (WN) process:*

$$z \sim WN(0, \sigma) \tag{1}$$

According to condition (1) of the Wold's theorem, if the fitted model belongs to the set of "correct" functional classes, the residuals should behave like a white noise process with zero mean and constant variance. Hypotheses whose residuals do not comply with condition (1) may be rejected using specific statistical tests that check randomness. The Ljung-Box test [5] is one of such tests and the one implemented in our system.

The null hypothesis of the Ljung-Box test is a strict white noise process. Thus, residuals are independent and identically distributed (i.i.d.). According to the definition of statistical independence [5], residuals are incompressible. Muggleton and Srinivasan [6], have also proposed to check noise incompressibility for evaluating hypothesis significance but in the context of classification problems.

Other statistical tests may be incorporated to check our assumptions regarding error structure, like tests for normality. The use of residuals for model assessment is a very general method which apply to many situations.

3.1 Relationship with Model Selection

The RMSE and hypothesis evaluation methods that are complexity aware do not guarantee that the model fits data well. For example, a model may have

the smallest overall RMSE of the candidate set but its residuals may present an increasing trend on the amplitude, or any other related curvature pattern that indicates lack-of-fit of the hypothesis and non independence of residual process. Thus, the Model Validation step introduced on the search process is of fundamental importance to assess the goodness-of-fit before proceeding to the choice of the best hypothesis because its scope is orthogonal to the model selection criterium.

3.2 Limitations discussion

The usage of Model Validation tests to restrict the set of acceptable hypothesis is supported by the White theorem [7], which briefly states that for a fixed set of models and a battery of validation tests, as the sample size tends to infinity and increasingly smaller test sizes are employed, the test battery will with probability 1 select the best model. The results of White theorem are asymptotic, which is not enough to assure good results in the sample sizes typically used in machine learning, nevertheless we have found good results in the empirical evaluation.

4 Experimental Evaluation

This section presents empirical evidence for the proposals made in this paper. We compare our approach results against other methodologies including the same ILP system without the modifications proposed in this paper.

4.1 Datasets

All experiments use the following datasets: Canada's Industrial Production Index [8]; USA Unemployment rate [9]; ECG of a patient with sleep apnea [10] and; VBR Traffic of an MPEG video [11].

The datasets consist of facts that relate time with the output variable. The time is expressed in discrete intervals and the output is a real-valued variable.

4.2 Benchmark models

This study illustrates the usage of an ILP system inducing switching models for non-linear time series. We have compared the results with structural, seasonal, linear and nonlinear time series models.

The theory induced by IndLog with model validation (denoted by IndLog*) is compared against the following models: IndLog without model validation; Auto-Regressive Integrated Moving Average (ARIMA); Threshold Auto-Regressive (TAR); Markov Switching Autoregressive (MSA); Autoregressive model with multiple structural Changes (MSC); Self-Excited Threshold Auto-Regressive (SETAR); Markov Switching regime dependent Intercepts Autoregressive parameters and (H)variances(MSIAH); Markov Switching regime dependent Means and (H)variances (MSMH); Bivariate Auto-Regressive models (Bivariate AR)

and; Radial Basis Functions Networks with structure optimization (RBFN). A detailed description of teh modesl may be found in the papers referenced in the previous subsection (datasets).

4.3 Learned task description

We propose to learn a modified class of the TAR [5] model.

The main difference regarding the original TAR structure is that instead of a single time delay value D, we have one D for each sub-region. The task to be learned consists of estimating: (i) the number of parameters p of each AR sub-model; (ii) The time delay D of each sub-model and; (iii) The thresholds that bound each subregion R_m and R_{m+1}. The generated clauses (hypothesis) are of the form: $\text{timeseries}(T, X) \leftarrow \text{inInterval}(R_m, T, D, R_{m+1}), \text{armodel}(T, P, X)$.

The ILP system Background Knowledge was provided with 12 AR(p) models, p ranging from 1 to 12; three model archetypes consisting of two, three and four amplitude intervals with 12 delay choices, D ranging from 1 to 12. The interval amplitude is calculated dividing the peak-to-peak value of each series by the number of intervals used in the archetype. We have configured the model validation step with Ljung-Box test with a p-value set to 0.75 in all datasets.

4.4 Results Summary and Discussion

This Section presents the results obtained for each dataset of Subsection 4.1. These datasets were studied in several papers, using different classes of models. Thus, all datasets in Table 1 have an AR model that may be used as a reference across datasets.

Table 1. Summary of results of the Relative RMSE of the ILP algorithm and other benchmark models for the selected datasets

Algorithm	Unempl	Prod	VBR	ECG
IndLog*	0.92	0.87	0.94	0.85
IndLog	1.11	1.04	0.96	0.93
AR	1.04	0.98	0.94	0.97
SARIMA	1.00	1.00	-	-
MSA	1.19	-	-	-
MSC	-	1.00	-	-
MSMH	-	0.98	-	-
MSIAH	-	1.20	-	-
SETAR	-	1.19	-	-
TAR	1.00	-	1.00	-
RBFN	-	-	-	1.00
Bivariate AR	1.20	-	-	-
Benchmark	1.59E-1	4.44E-3	12.9E3	4.53

IndLog* consistently induced models with best forecasting performance in the test sets of all datasets. This allows us to conclude that the proposed modifications to the basic ILP search process, makes an ILP system suited for learning time-series models and discovering new model structures.

In all datasets, the ILP system was configured to allow the induction of the TAR model proposed in the cited papers. However, all models induced by the ILP system, used different D values in the constraint that selects each sub-region. This is perhaps an indicator of the superior flexibility of this structure to capture implicit seasonality and or complex switching behavior in those time series.

5 Conclusions

In this paper we presented preliminary results of using a statistical-based criterium for hypotheses acceptance, which we designate model validation, that assess the *goodness-of-fit* of the induced hypotheses in numerical domains.

Our proposals were incorporated in the IndLog [12] ILP system and evaluated on time series modelling problems. The ILP results were better than other statistics-based time series prediction methods. The ILP system discovered a new switching model based on the possibility of varying the delay on the activation rule of each sub-model of a TAR model.

References

1. A. Srinivasan and R. Camacho. Numerical reasoning with an ILP system capable of lazy evaluation and customised search. *J. Logic Prog.*, 40(2-3):185–213, 1999.
2. Hendrik Blockeel, Luc De Raedt, and Jan Ramon. Top-down induction of clustering trees. In J. Shavlik, editor, *15th ICML*, pages 55–63. Morgan Kaufmann, 1998.
3. S. Dzeroski and L. Todorovski. Declarative bias in equation discovery. In *14th ICML*, pages 376–384, USA, 1997.
4. A Karalič. Employing linear regression in regression tree leaves. In *ECAI*, pages 440–441, Chichester, 1992. John Wiley & Sons.
5. H. Tong. *Nonlinear time series, a dynamical system approach*. Claredon press, Oxford, UK, 1st edition edition, 1990.
6. S. Muggleton, A. Srinivasan, and M. Bain. Compression, significance and accuracy. In D. et al. Sleeman, editor, *ML92*, pages 338–347. Morgan Kauffman, 1992.
7. K. Hoover and S. Perez. Data mining reconsidered: encompassing and the general-to-specific approach to specification search. *Econometrics J.*, 93(2):167–191, 1999.
8. B. Silvertovs and D. Dijk. Forecasting industrial production with linear, nonlinear and structural change models. Technical report, Erasmus University, 2003.
9. R. Tsay l. Montgomery, V. Zarnowitz and G. Tiao. Forecasting the u.s. unemployment rate. *JASA*, 93:478–493, 1998.
10. M. et al Kreutz. Structure optimization of density estimation models applied to regression problems. In *Proc. of the 7th Int. Workshop on AI and Statistics*, 1999.
11. B. Jang and C. Thomson. Threshold autoregressive models for vbr mpeg video traces. In *IEEE INFOCOM*, pages 209–216, USA, 1998.
12. R. Camacho. *Inducing Models of Human Control Skills using Machine Learning Algorithms*. PhD thesis, Universidade do Porto, July 2000.