Improving Numerical Reasoning Capabilities of Inductive Logic Programming Systems

Alexessander Alves, Rui Camacho and Eugenio 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 22 508 184 fax: +351 22 508 1443

Abstract. Inductive Logic Programming (ILP) systems have been largely applied to classification problems with a considerable success. The use of ILP systems in problems requiring numerical reasoning capabilities has been far less successful. Current systems have very limited numerical reasoning capabilities, which limits the range of domains where the ILP paradigm may be applied.

This paper proposes improvements in numerical reasoning capabilities of ILP systems. It proposes the use of statistical-based techniques like Model Validation and Model Selection to improve noise handling and it introduces a new search stopping criterium inspired in the PAC learning framework.

We have found these extensions essential to improve on results over statistical-based algorithms for time series forecasting used in the empirical evaluation study.

1 Introduction

Inductive Logic Programming (ILP) [1] has achieved considerable success in domains like biochemistry [2], language processing [3], environment monitoring [4]. The success of those applications are mainly due to the intelligibility of the models induced. Those models are expressed in the powerful language of first order clausal logic. In the domains just mentioned, the background knowledge is mainly of a relational nature. Theoretically there is no impediment of using whatever knowledge is useful for the induction of a theory. For some applications it would be quite useful to include as background knowledge methods and algorithms of a numerical nature. Such an ILP system would be able to harmoniously combine relations with "numerical methods" in the same *model*. A proper approach to deal with numerical domains would therefore extend the applicability of ILP systems. It would also pave the way for more sophisticated applications, like discovering new time series model structures.

Current ILP approaches [5] to numerical domains usually carry out a search through the model (hypothesis) space looking for a minimal value of a cost function like the Root Mean Square Error (RMSE). Systems like TILDE [6] are of that kind. One problem with the minimisation of RMSE in noisy domains is that the models tend to be brittle. The error is small when covering a small number of examples. The end result is a large set of clauses to cover the complete set of examples. This is a drawback on the intelligibility of ILP induced models. This aspect is also an obstacle to the induction of a *numerical theory*, since we end up with small locally fitted *sub-models*, that may not correspond to the overall structure of the underlying process that generated data.

In this paper we propose improvements on the numerical reasoning capabilities of ILP systems by adopting statistical-based noise handling techniques such as: (i) model validation and; (ii) model selection.

We also propose a new stopping criterion inspired on the PAC [7] framework.

The rest of the paper is organised as follows. Section 2 identifies the steps of a basic ILP algorithm that are subject to improvements proposed in this paper. The proposals for Model Validation are discussed in Section 3. In Section 4 we propose the the stopping criterion. The proposals for Model Selection are discussed in Section 5. Section 6 presents the experimental findings. The related work is discussed in Section 7. Finally, in Section 8 we draw the conclusions.

2 Search Improvements

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 classification tasks is to score a hypothesis by measuring its coverage. That is, the number of examples it explaines. In numerical domains it is common to use the RMSE or Mean Absolute Error (MAE) as a score measure. Algorithm 1 presents an overview of the procedure.

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

```
 1: repeat
 2:     Initialize K
 3:     repeat
 4:         $h_i \leftarrow$ synthesize an hypothesis
 5:         accept an hypothesis (Model Validation)
 6:         if Stopping Criterion satisfied then
 7:             $K \leftarrow K - 1$
 8:             Update best hypothesis (Model Selection)
 9:         end if
10:     until $K = 0 \vee h_i = \varnothing$
11:     Remove explained examples
12: until “All” examples explained
```

We propose an improvement to step 5 where an hypothesis is checked if it is a satisfactory approximation of the underlying process that generated data. We propose the use of statistical tests in that model validation step. Step 8 is improved avoiding the overfitting problem, which is manifest in the fragmented structure of the induced theories, using a model selection criterium. Our proposal for step 6 is inspired on the PAC [7] framework.

We use the terms hypothesis, model and theory with the following meaning in this paper. An hypothesis (clause) is a conjecture after a specific observation and before any empirical evaluation has been performed. A model is an hypothesis that has at least limited validity for predicting new observations. A model is an hypothesis that has passed the model validation tests. A Theory is a set of hypotheses that have been confirmed through empirical evaluation.

3 Model Validation

In most applications, the true nature of the model is unknown, therefore, it is 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 both in 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, defined as follows.

Definition 1 (Hypothesis Residuals). *The residuals from an induced hypothesis h_j are the differences between the responses observed at each combination values of the explanatory variables and the corresponding prediction of the response computed using the induced hypothesis. Mathematically, the definition of the residual z_i for the i^{th} observation in the data set is written, $z_i = y_i - h_j(\overrightarrow{x_i})$, where y_i denotes the i^{th} response in the data set and $\overrightarrow{x_i}$ represents the list of explanatory variables at the corresponding values found in the i^{th} observation in the data set. Therefore, residuals are the random process formed from the differences between the observed and predicted values of a variable.*

As a consequence of the Wold's theorem the behavior of the residuals may be used to check the adequacy of the fitted model.

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, defined below, is one of such tests.

Definition 2 (Ljung-Box Test). *The Ljung-Box test is based on the autocorrelation function. However, instead of testing randomness at each distinct lag, it tests the "overall" randomness based on a number of lags. The null hypothesis is the data are random. The test statistic is:* $Q_{LB} = n(n+2)\sum_{j=1}^{h} \frac{r(j)^2}{(n-j)}$. *Where* n *is the sample size,* $r(j)$ *is the autocorrelation at lag* j*, and* h *is the number of lags being tested. The Significance Level is* α*. The hypothesis of randomness is rejected if:* $Q_{LB} > \chi^2_{((1-\alpha);\,h)}$ *where* $\chi^2_{((1-\alpha);\,h)}$ *is the percent point function of the chi-square distribution.*

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, namely condition (2), residuals are incompressible. Muggleton and Srinivasan [8], 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.

Definition 3 (Statistical Independence). *Let* $x_1, x_2, \ldots, x_i$ *be a sequence of random variables. The variables* x_i *are statistically (mutually) independent if:*

$$\mathrm{E}\left[g_1(x_i)g_2(x_j)\right] - \mathrm{E}\left[g_1(x_i)\right]\mathrm{E}\left[g_2(x_j)\right] = 0 \quad \forall_{i \neq j}, \quad \forall_{g_1, g_2} \tag{2}$$

Notice that this is much stronger condition then uncorrelation: $\mathrm{E}\left[X_i X_j\right] - \mathrm{E}\left[X_i\right]\mathrm{E}\left[X_j\right] = 0 \quad \forall_{i \neq j}$. Condition (2) means that there is no function that captures the relationship between these random variables.

The formulation presented in this paper considers that only noise is statistically independent, which is an assumption based on the Wold's theorem and verifiable by the Ljung-Box Test.

4 Stopping Criterium

The stopping criterium derived is inspired on the PAC [7] method to evaluate learning: $P(|z| > \epsilon) < \delta$. The stopping criterium stops the search whenever the probability of the error be greater than the accuracy (ϵ) is less than the confidence interval (δ). Different degrees of "goodness" will correspond to different values of ϵ and δ.

In this section we propose to calculate the bound, δ, for any unknown distribution. Theorem 2, proves the existence of the bound, δ, for a single clause (hypothesis) and provides a procedure to calculate the error probability for a given accuracy level. Corollary 1, generalises the bound on the error probability to a multi-clausal theory. Both theorems rely on the convergence in probability mode, defined as follows.

Definition 4 (Convergence in probability). *Let* $x_1, x_2, \ldots, x_n$ *be a sequence of random variables. We say that* x_n *converges in probability to another random variable* x*, i.e.* $x_n \xrightarrow{P} x$*, if for any* $\epsilon > 0$*,* $P(|x_n - x| > \epsilon) \to 0 \quad as \quad n \to \infty$

Theorem 2 (Bounding Error Probability of an Hypothesis). *Let z be the residuals from the hypothesis h_i. Assume z is independent and identically distributed (i.i.d.) with distribution variance σ^2. Then the probability of the error being greater then ϵ is bounded by:*

$$P(|z| > \epsilon \mid h_i) < \delta, \quad \delta = \frac{\sigma^2}{\epsilon^2} \tag{3}$$

Proof: Let the residuals $z_1, z_2, \ldots, z_n$ be a sequence of i.i.d. random variables each with finite mean μ and σ. if $\bar{z} = (z_1 + \ldots + z_n)/n$ is the average of $z_1, z_2, \ldots, z_n$, then, it follows from the week law of large numbers [9] that:

$$\bar{z} \xrightarrow{P} \mu \tag{4}$$

Let the sample variance be $S_n = \frac{1}{n}\sum_{j=1}^{n}(z_j - \bar{z})^2 = \frac{1}{n}\sum_{j=1}^{n} z_j^2 - \bar{z}^2$, where $\bar{z}$ is the sample average. It follows from the Slutski's lemma [9] that:

$$S_n \xrightarrow{P} \sigma \tag{5}$$

Assuming the residuals z of the hypothesis h_i pass the null hypothesis of the Ljung-Box test, then they will comply with a strict white noise process with zero mean and finite variance, yielding thereby:

$$\mu = 0, \qquad \sigma < \infty \tag{6}$$

Following Conditions (4) and (5), each observation may be considered drawn from the same ensemble distribution. Thus, the sample mean and variance of the joint distribution converge to the ensemble mean and variance. Moreover, condition (6) states that both values are finite and, therefore, for all $\epsilon > 0$, the Chebishev's inequality bounds the probability of the residuals value, z, being greater then ϵ to:

$$P(|z| > \epsilon \mid h_i) < \frac{\sigma^2}{\epsilon^2} \tag{7}$$

■

Corollary 1 (Bounding Error Probability of a Theory). *Let H be a set of hypothesis (clauses) that describes a given theory T. Assume:*

$$P(|z| > \epsilon \mid h_i) < \delta, \quad \forall_{h_i \in H} \tag{8}$$

then, for theory T, the probability of the error, z, being greater than ϵ, is also bounded by $P(|z| > \epsilon) < \delta$.

We recall that just one clause is activated at each time thus all clauses of a theory are mutually exclusive regarding example coverage, i.e.

$$h_i \cap h_j = \emptyset \quad \forall_{i \neq j} \tag{9}$$

We also recall that the prior probability of h_i, $P(h_i)$ may be estimated calculating the frequency of h_i on the training set and dividing it by the coverage of the theory. Because the sum of the frequencies of all hypotheses is equal to the theory coverage, then

$$\sum_{\forall_{h_i \in H}} P(h_i) = 1 \tag{10}$$

Proof: Let conditions (9) and (10) hold, then it follows from the *total probability theorem* that:

$$P(|z| > \epsilon) = \sum_{\forall_{h_i} \in H} P(|z| > \epsilon \mid h_i) P(h_i). \tag{11}$$

Let condition (8) hold, then we may substitute $P(|z| > \epsilon \mid h_i)$ by δ in equation (11), yielding thereby: $P(|z| > \epsilon) < \delta \sum_{\forall_{h_i \in H}} P(h_i)$. Since $\sum_{\forall_{h_i \in H}} P(h_i) = 1$ and $P(|z| > \epsilon \mid h_i) < \delta, \quad \forall_{h_i \in H}$, then:

$$P(|z| > \epsilon) < \delta \tag{12}$$

■

5 Model Selection

The evaluation of conjectured hypotheses is central to the search process in ILP. Given a set of hypothesis of the underlying process that generated data, we which to select the one that best approximates the "true" process. The process of evaluating candidate hypothesis is termed *model selection*.

A simple approach to model selection is to select the hypothesis that gives the most accurate description of data. For example, select the hypothesis that minimises RMSE. However, model selection is disturbed by the presence of noise in data, leading to the problem of over fitting. Thus, an hypothesis with larger number of adjusted parameters has more flexibility to capture complex structures in data but also to fit noise. Hence, any criterium for model selection should establish a trade-off between descriptive accuracy and hypothesis complexity.

5.1 Hypothesis Complexity

Defining a theoretically well-justified measure of model complexity is a central issue in model selection that is yet to be fully understood. In Machine Learning, some authors have advanced their own definition of complexity. Dzerovski [10], proposes a complexity measure based on the length of a grammar sentence in the Lagramge system. Muggleton [11] proposes a complexity measure based on the number of bits necessary to encode an hypothesis.

Both complexity measures are sensitive to the hypothesis functional form. This is clear since both penalises each literal added. The functional form is not a good approximation to measure the complexity of a real-valued hypothesis, since any real-valued function can be accurately approximated using a single function class. This follows directly from Approximation Theory. An example is the Kolmogorov's superposition theorems.

Theorem 3 (Kolmogorov superposition theorem). *Any continuous multidimensional function $f(x_1, \ldots, x_m)$, can be represented as the sum of $m + 1$ functions. These functions are called universal functions because depend only on the dimensionality m and not in the functional form of f.*

Following theorem 3, the sum of universal functions is proportional to the dimensionality m. This highlights the role of dimensionality on a definition of hypothesis complexity. A few arguments on computational complexity and estimation theory also support this claim. Since the machine learning algorithm is given a finite dataset, models with fewer adjusted parameters will be easier to optimise since they will generically have fewer misleading local minima in the error surfaces associated with the estimation. They will be also less prone to the curse of dimensionality. They will require less computational time to manipulate. A model with fewer degrees of freedom generically will be less able to fit statistical artifacts in small data sets and will therefore be less prone to the so-called "generalisation error". Finally, several authors (Akaike [12]; Efron [13]; Ye [14]) proposed measures of model complexity which in general depend on the number of adjusted parameters. Consequentially, the adopted measure of complexity in this work is the number of adjusted parameters to data.

5.2 Model Selection Criteria

There are several model selection criteria suitable for the adopted measure of model complexity. Among these, we may find: (i) Akaike Information Criterium (AIC) [12], defined as $AIC = -2\ln(L) + 2k$; (ii) Akaike Information Criterium Corrected for small sample bias(AICC) [12], defined as $AICC = -2\ln(L) + 2k\frac{n}{n-k+1}$; (iii) Bayesian Information Criterium (BIC) [15], defined as $BIC = -\ln(L) + \ln(n)k$ and; (iv) the Minimum Description Length (MDL) [12], defined as $MDL = -\ln(L) + \frac{k}{2}\ln(n) + (\frac{k}{2}+1)\ln(k+1)$.

The estimation of an hypothesis likelihood function, L, with k adjusted parameters, requires a considerable computational effort and the assumption of prior distributions. In this context, the Gaussian distribution plays an important role in the characterisation of the noise, fundamentally due the central limit theorem. Assuming error is i.i.d. drawn from a Gaussian distribution then, the likelihood of an hypothesis given the data [12] is: $\ln(L) = -\frac{n}{2}(1 + \ln(2\pi) + \ln(\hat{\sigma_r}^2))$, where $\hat{\sigma}_r^2 = \frac{1}{n}\sum_{i=1}^{n} z_i^2$, and z are the residuals of the induced hypothesis.

Analytical model selection criteria like AIC and BIC are asymptotically equivalent to leave-one-out and leave-v-out cross-validation [16]. However, they have the advantage of being incorporated in the cost function.

When these hypotheses have different coverages, Box and Jenkins [17](pg. 201) suggests the normalisation of those criteria by the sample size. This approach has the advantage of indirectly biasing the search to favour hypothesis with higher coverage, and consequentially, theories with less clauses.

Other authors presented similar work in this area. Zelezni [18] derives a model selection criterium under similar assumptions that uses the Muggleton's complexity measure, which according to the adopted definition of complexity, is unsuitable for our purposes. It also requires the calculation of the “generality” function for each induced hypothesis. His formulation does not estimate the modal value of the likelihood, so the final equation includes the usually unknown nuisance parameter of the hypothesis, which somehow limits its practical use.

5.3 Choosing a Model Selection Criterium

The adopted model selection criteria have different characteristics, thus, it is essential to clarify their application conditions to numerical problems in ILP.

The use of AIC is recommendable if the data generating function is not in any of the candidate hypotheses and if the number of models of the same dimension does not grow very fast in dimension, then the average squared error of the selected model by AIC is asymptotically equivalent to the smallest possible one offered by the candidate models [16]. Otherwise, AIC cannot be asymptotically optimal, increasing model complexity as more data is supplied [15].

The use of BIC and other dimension consistent criteria like MDL is advisable if the correct models are among the candidate hypothesis, then the probability of selecting the true model by BIC approaches 1 as $n \to \infty$. Otherwise, BIC has a bias for choosing oversimplified models [16].

6 Experimental Evaluation

This section presents empirical evidence for the proposals made in this paper. We propose an experiment that illustrates the usage of an ILP system on scientific discovery tasks. In that sense, this experiment has been inspired in Colton and Muggleton [19] application of an ILP system to mathematical discovery. The experiment consists of learning a model for time series prediction using an ILP system. The model extends a previously existent one, by adding extra degrees of freedom that will be estimated in run-time by the ILP system. Although this experiment reports on learning multiple clause theories, it is related with Zelezni’s [18] work because it also learns a numeric function.

6.1 Datasets

Canada’s Industrial Production Index [20]; USA Unemployment rate [21]; ECG of a patient with sleep apnea [22] and; VBR Traffic of an MPEG video [23]. These datasets consist of facts that relate time with an output variable. The time is expressed in discrete intervals and the output is a real-valued variable. The mode declaration for the head literal is of the form: timeseries(+Time, −Output).

6.2 Benchmark models

In this experiment we compared theories induced by the raw IndLog system with the following models: IndLog with Model Validation and Model Selection activated (IndLog$^{\mathrm{MVS}}$); 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 (RBFN)

All models are described in the papers referred in the datasets section. In all experiments the statistics used to compare the models is the Root Mean Square Error (RMSE). All time series models use forecasting lead time of one period.

6.3 Learning Task Description

The experiment's goal is to learn a modified class of the TAR [24] model.

Definition 5 (TAR Model). *The TAR model is a nonlinear time-series process composed of linear AR(p) sub-models. Each amplitude switched AR process is constructed for a specific amplitude subregion. The AR model to be used at time n is determined by the amplitude $x(n-D)$ where D denotes a time-delay. The AR model for sub-region m is activated if the following constraint is true: $R_m < x(n-D) < R_{m+1}$. The variable x is the time-series observed, m is the index denoting the sub-region, R_m denotes the threshold amplitude of region m.*

The main difference from the original TAR structure is that instead of a single D value, we have one D for each sub-region.

The learning task 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 induced clauses are of the kind: $\text{timeseries}(T, X) \leftarrow \text{inInterval}(R_m, T, D, R_{m+1}), \text{armodel}(T, P, X)$.

6.4 Results Summary and Discussion

This section presents the results obtained for each dataset of Section 6.1. Those datasets were studied in several papers, using different classes of models. Thus, all time series datasets in table 1 have an AR model that may be used as a reference across datasets. The recall number for the Unemployment, Production, VBR Traffic, and ECG datasets are respectively: 100%, 96%, 94%, and 78%.

The ILP system consistently induced models with best forecasting performance on all datasets studied. This allow us to conclude that the proposed modifications to the basic ILP search process, makes an ILP system suited for discovering new time series models.

7 Related Work

Other approaches to the task of learning numerical relationships may be found in the ILP literature. FORS [25] integrates feature construction into linear regression modelling. The ILP system IndLog [26] presented mechanisms for coping

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

Model	Unemployment	Production	VBR Traffic	ECG
IndLog^{MVS}	0.91	0.85	0.93	0.82
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 RMSE	1.59E-1	4.44E-3	12.93E3	4.53

with large number of examples that include noisy numerical data without negative examples and the capability to adjust model parameters at induction time. Equation discovery systems like LAGRAMGE [27] allow the user to specify the space of possible equations using a context-free grammar. TILDE [6] has the capability of performing regression-like tasks.

8 Conclusions

In this paper we have proposed improvements in the numerical reasoning capabilities of ILP systems. The improvements proposed are: model validation; model selection criteria and; a stopping criterium.

Our proposals were incorporated in the IndLog [26] 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.

The proposals made for model validation, model selection and for measuring the learning performance can be generalised to other machine learning techniques dealing with numerical reasoning.

References

1. S. Muggleton. Inductive logic programming. *New Generation Computing*, 8(4):295–317, 1991.
2. R.D. King, S. Muggleton, A. Srinivasan, and M. Sternberg. Structure-activity relationships derived by machine learning. *Proceedings of the National Academy of Sciences*, 93:438–442, 1996.

3. J. Cussens. Part-of-speech tagging using progol. In *In* Proc. of the 7th Int. Workshop on Inductive Logic Programming, pages 93–108. Springer, 1997.
4. S. Džeroski, L. Dehaspe, B. Ruck, and W. Walley. Classification of river water quality data using machine learning. In *5th Int. Conference on the Development and Application of Computer Techniques to Environmental Studies*, 1994.
5. 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.
6. Hendrik Blockeel, Luc De Raedt, and Jan Ramon. Top-down induction of clustering trees. In J. Shavlik, editor, *Proceedings of the 15th International Conference on Machine Learning*, pages 55–63. Morgan Kaufmann, 1998.
7. L. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.
8. S. Muggleton, A. Srinivasan, and M. Bain. Compression, significance and accuracy. In D. et al. Sleeman, editor, *ML92*, pages 338–347. Morgan Kauffman, 1992.
9. P. Brockwell and R. Davis. *Time series: theory and methods.* Springer, N.Y., 1991.
10. S. Dzeroski and L. Todorovski. Discovering dynamics. In *10th ICML*, pages 27–29, Massachusetts, USA, 1993.
11. S. Muggleton. Learning from positive data. In S. Muggleton, editor, *ILP96*, volume 1314 of *LNAI*, pages 358–376. Springer, 1996.
12. K. Burnham and D. Anderson. *Model Selection and Multimodel Inference.* Springer, New York, 2002.
13. B. Efron. How biased is the apparent error rate of a prediction rule? *JASA*, 81:461–470, 1986.
14. J. Ye. On measuring and correcting the effects of data mining and model selection. *JASA*, pages 120–131, 1998.
15. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.
16. J Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264, 1997.
17. Jenkins Box and Reinsel. *Time Series Analysis, Forecasting and Control.* Prentice Hall, Englewood Cliffs, N.J., USA, 3rd edition edition, 1994.
18. Filip Zelezny. Learning functions from imperfect positive data. In *International Workshop on Inductive Logic Programming*, pages 248–260, 2001.
19. S. Colton and S. Muggleton. ILP for mathematical discovery. In T. Horváth and A. Yamamoto, editors, *ILP03*, volume 2835 of *LNAI*, pages 93–111. Springer, 2003.
20. B. Silvertovs and D. Dijk. Forecasting industrial production with linear, nonlinear and structural change models. Technical report, Erasmus University, 2003.
21. R. Tsay l. Montgomery, V. Zarnowitz and G. Tiao. Forecasting the u.s. unemployment rate. *JASA*, 93:478–493, 1998.
22. 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.
23. B. Jang and C. Thomson. Threshold autoregressive models for vbr mpeg video traces. In *IEEE INFOCOM*, pages 209–216, USA, 1998.
24. H. Tong. *Nonlinear time series, a dynamical system approach.* Claredon press, Oxford, UK, 1st edition edition, 1990.
25. A. Karalic and I. Bratko. First order regression. *Machine Learning*, 26:147–176, 1997.
26. R. Camacho. *Inducing Models of Human Control Skills using Machine Learning Algorithms.* PhD thesis, Universidade do Porto, July 2000.
27. S. Dzeroski and L. Todorovski. Declarative bias in equation discovery. In *14th ICML*, pages 376–384, USA, 1997.