Approximate Identification of Automata

Brian R. Gaines
Department of Electrical Engineering Science
University of Essex
Colchester, Essex, England

A technique is described for the identification of probabilistic and other non-deterministic automata from sequences of their input/output behaviour. For a given number of states the models obtained are optimal in well defined senses, one related to least-mean-square approximation and the other to Shannon entropy. Practical and theoretical investigations of the technique are outlined.

1 Introduction

The use of the Nerode equivalence over regular sets to derive a minimum-state automaton that realises the input/output behaviour of some discrete system is well established, and various practical implementations have been developed (Fu and Booth, 1975), as have rigorous mathematical foundations (Goguen, 1973). The combinatorial difficulty of determining the equivalence and the corresponding computational problem are well known and minimum-time algorithms have been developed (Hopcroft, 1971). However, a more fundamental limitation to the use of the technique for system identification has been described recently: the Nerode equivalence gives exactly the correct structure in its original context of deterministic systems, but the introduction of the slightest observation noise, or acausality in the observed system, leads to a meaningless structure whose member of states grows proportionally with the length of observations (Gaines, 1976). This letter describes a new approach to behaviour→structure transformations that yields the same results as the Nerode equivalence for deterministic systems, but also identifies non-deterministic and probabilistic systems.

2 Problem Statement

From an observed sequence of behaviour, e.g. *bcccdeecf,* where *b is* the first observation and *f* the last, infer the 'best' structure for an automaton that might generate it. More formally, an *observed behaviour* is a member *of* the free monoid D^* generated by some set *of descriptors D of* which some subset *I* may be designated as *inputs* (which need not be predicted); some subset *T* as *terminators (such* that the string before them may not be used to predict the string after them) and the remainder as *outputs.* However, neither *I* nor *T* need be defined and the algorithm will determine them (at the cost of additional computation). For the purposes of modelling, 'null' inputs and outputs are assumed to be inserted in any member of D^* if necessary to give an interleaved sequence.

A model space $<M, C, P>$ for the problem consists of an allowed set of models *M*, a function $C: M \rightarrow R^+$, from models to the positive reals, measuring the *cost* of a model and a function $P: M \times D^* \rightarrow R^+$, measuring the *poorness of fit* of a model to an observed behaviour. An *admissible subspace* for a given $u \in D^*$ *is* determined by $M(u) \subset M$, the subset of nonimprovable models such that if $m \in M(u)$ there is no $m' \in M$ such that $P(m', u) < P(m, u)$ and $C(m') \leq C(m)$. For example, *M* might be the set of irreducible, deterministic Mealy machines with a specified initial state, *C(m)* might be the number or of states of $m \in M$ and *P(m, u)* might be 0 if *m* generates the input/output sequence $u \in D^*$ exactly, and 1 otherwise (where m starts in the specified initial state, receives the input sequence embedded in *u* and returns to the initial state at any terminator in *u*). This corresponds to the standard deterministic case in which the Nerode equivalence may be used to derive an admissible subspace consisting in fact of a set of isomorphic minimal-state automata.

The binary nature of *P* in this case corresponds to there being a well defined correct solution given the allowed class of models. However, Gaines (1976) shows that the model obtained is not meaningful if the observed behaviour has been generated by a non-deterministic system. *M* needs to be widened to allow for approximate models, and the evaluation *P* needs to reflect the degree of approximation.

3 Approximate Models

Let *M* now be the set of probabilistic Mealy machines that are observable in the sense that, even though the next state can only be predicted probabilistically, each possible transition is associated with a distinct output, so that the state after a transition may be determined from the output. The inputs and terminators in a sequence u will again drive it through a well defined state trajectory, but now it is not expected to exactly match each output in a, only to predict it. *C(m)* is again the number of states in *m,* but *P(m, a) is* a measure of the poorness of prediction. One possible measure P_e is the total errors when *m is* used to predict only the most likely output. However, this is unnecessarily coarse, since *m* has available a vector of probabilities over possible outputs, and, for example, c occurring when the prediction of *b, c, d* is (0 6, 0.4, 0) is clearly much better than when it is (0.6, 0, 0.4).

The literature on subjective probability provides bases for measures with better discrimination than P_e Finetti (1972) has shown that, if the actual event is represented by a vector with a 1 for the output which occurs and 0 for the others, the distance between the prediction and occurrence is a suitable measure of poorness of performance. The total of the squares of the Cartesian distances between the predictions of a model m and the occurrences of outputs in a sequence u will be denoted $P_s(m, a)$. Finetti shows that if the occurrences form an ergodic stochastic process, a predictor minimising its mean-square distance will come to predict the actual probabilities of occurrences. Other performance measures with this property have been characterised (Winkler and Murphy, 1968) and a particularly interesting one is that which only takes into account the component of the prediction for the event which occurs allocating a poorness estimate of the logarithm of the predicted 'probability' to it. The total of the logarithms to base 2 of the probabilities predicted by a model *m* of the outputs in a sequence u will be denoted $P_l(m, u)$.

4 Implementation

Algorithms to compute the admissible subspace of either Mealy or Moore probabilistic observable automata for arbitrary u have been implemented and tested with a wide variety of source material. The algorithm is basically a search over non-deterministic automata—the optimality condition mentioned for the measures implies that the 'transition probabilities' may be filled in from the measured transition frequencies of these automata. A recursive algorithm constrained by the sequence to generate only possible models is used, but otherwise the search is exhaustive. The output from the program is a listing of admissible automata for increasing *C(m)* (number of states), together with values of P_e, P_s and P_l for each automaton.

5 Outline of Results

Practically, the technique is limited by the amount of computation required. and in a typical run (10 h on a l microsec-cycle-time 16 bit-word minicomputer) models of up to 10 states only may be derived. However. examples of sequences from deterministic, probabilistic and asynchronous machines together with human problem-solving data have been successfully analysed. Fig. 1 is a plot of $P_l(m, u)$ against $C(u)$ from the analysis of a 100-element binary sequence generated by a stochastic learning environment studied by Andreae (1974), and also shows the beginning of the observed behaviour and the 2-, and 4-, state admissible models generated. The 4-state model accurately identifies Andreae's original stochastic environment, and a pronounced drop in P_l may be noted at 4 states. P_l becomes 0 for the 91-state model, in line with the results of (Gaines, 1976).

Theoretically, it may be shown that:-

(a) if at is an ergodic stochastic sequence of outputs generated by an n-state automaton the expected value of $P_s(n, u)$ is a measure of the entropy of the sequence $[\Sigma p(1-p)]$;

(b) under the same conditions the expected value of $P_l(n, u)$ is minus the Shannon entropy of the sequence $[\Sigma p(\log_2 p)]$;

(c) under the same conditions, if the generating probabilities are not 0 or 1 the expected value of $C(m)$ at which $P(m, u) = 0$ is at least half the length of u (Gaines, 1976);

(d) no improvement can be gained by considering non-observable automata as models.

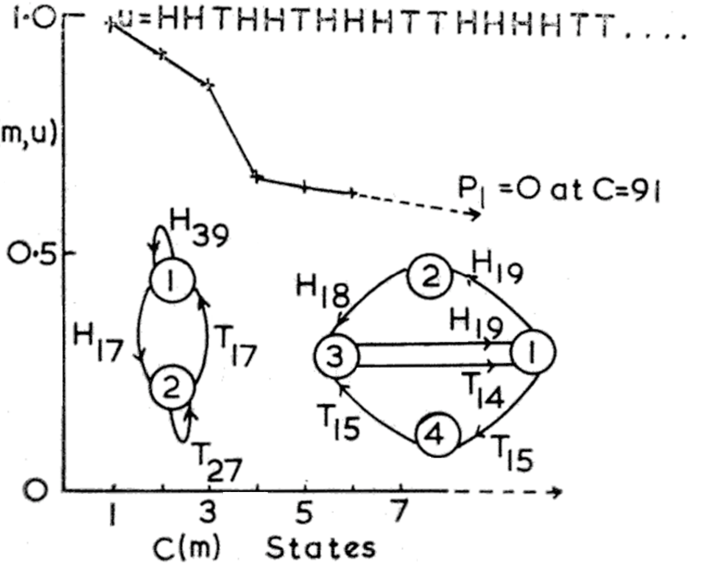

Figure 1 Results of modelling sequence u (initial part shown)

Graph of poorness of fit P_l against number of states C for admissible models: 2-state and 4-state admissible models generated (the subscripts to the labels of the transitions are the number of times the transition occurs in u)

6 Conclusions

A technique has been given for the determination of automaton structures from a sequence of behaviour which is 'best' in some rigorous and meaningful sense. It overcomes the problems associated with extending deterministic techniques to the non-deterministic case, accurately identifying probabilistic automata and providing a pseudo-probabilistic model for other non-deterministic sources. Practically, its use as a system identification technique is feasible in simple cases, but is limited by computation time to models with up to 10 states. Nevertheless it provides a normative technique against which faster heuristic methods may be judged. Theoretically, it gives a finite-state approach to the modelling and measurement of computational complexity of individual sequences (Kolmogorov, 1968). It is also a base-level minimum-assumption method, against which the savings of assumptions of deterministic source, known maximum number of states, linearity etc., may be evaluated.

References

Andreae, J.H. (1974). PURR-PUSS: purposeful, unprimed, rewardable robot. University of Canterbury. Man-machine Studies Electrical Engineering Report 24.

De Finetti, B. (1972). **Probability, Induction and Statistics: The Art of Guessing**. London, Wiley.

Fu, K.S. and Booth, T.L. (1975). Grammatical inference: Introduction and survey - Part I. **IEEE Transactions on Systems, Man & Cybernetics SMC-5** 95-111.

Gaines, B.R. (1976). On the complexity of causal models. **IEEE Transactions on Systems, Man & Cybernetics SMC-6**(1) 56-59.

Goguen, J.A. (1973). Realization is universal. **Mathematical Systems Theory 6** 359-374.

Hopcroft, J. (1971). An n log n algorithm for minimizing states in a finite automaton. Kohavi, Z. and Paz, A., Ed. **Theory of Machines and Computations**. pp.189-196. New York, Academic Press.

Kolmogorov, A.N. (1968). Logical basis for information theory and probability. **IEEE Transactions on Information Theory IT-14** 662-664.

Winkler, R.C. and Murphy, A.H. (1968). Good probability assessors. **Journal of Applied Meteorology 7** 751-758.