Fuzzy Sets and Systems 2 (1979) 15–24.

SEQUENTIAL FUZZY SYSTEM IDENTIFICATION

Brian R. GAINES

Man-Machine Systems Laboratory, University of Essex, Colchester, U.K.

The problem of deriving thc structure of a non-deterministic system from its behaviour is a difficult one even when that behaviour is itself well-defined. When the behaviour can be described only in fuzzy terms structural inference may appear virtually impossible. However, a rigorous formulation and solution of the problem for stochastic automata has recently been given [1] and, in this paper, the results are extended to *fuzzy stochastic automata and grammars.* The results obtained are of interest on a number of counts. (1) They are a further step towards an integrated 'theory of uncertainty'; (2) They give new insights into problems of inductive reasoning and processes of 'precisiation'; (3) They are algorithmic and have been embodied in a computer program that can be applied to the modelling of sequential fuzzy data; (4) They demonstrate that sequential fuzzy data may be modelled naturally in terms of 'possibility' vectors.

Key Words: Possibility, Identification, Automata, Grammars, Modelling, Complexity, Approximation.

1. Introduction

One of the system-theoretic problems that Lotfi Zadeh was studying in the decade before his seminal work on fuzzy systems was that of deriving the *structure* of a system from observations of its *behaviour.* In a 1956 paper [2] he coined the word *identification* as a generic term for the variety of forms of behaviour/structure inference problem then being studied. These varied widely, mainly in the forms of structure considered: from linear, continuous systems, through diverse weakenings of linearity and continuity, to general automata with no signal-space or state-space topologies or constraints.

The next 20 years have seen the development of computer systems for on-line control, and system identification has become a major area of research in its own right [3] generating a continuing series of major IFAC conferences concerned with that topic alone. Control-theoretic interest has naturally tended to concentrate on systems modelled as linear and continuous in their signal and state variables, and either continuous in time or uniformly sampled. However, the more general forms of model have also found applications, particularly in the study of biological [4–6] and human control systems [7].

Noam Chomsky's classic paper, Three models for the description of language [8], appeared in the same year as Zadeh's on identification. The link between automata, generative grammars, and natural languages that Chomsky proposed initiated a major area of research in modern linguistics, and the problem of how a child might *acquire* the grammatical structures through conversational experience led to Solomonoff's studies [9, 10] of *inductive grammatical inference.* Moore's classic paper, Gedanken experiments on sequential machines [11], that considered related problems for finite-state automata was a key stimulus in triggering off Zadeh's generalization of the behaviour/structure inference problem, and Solomonoff's work on inductive inference also fitted into this general framework.

Over the next 20 years grammatical inference became a major research topic in its own right [12]. The problem for finite-state automata or grammars was shown by Rabin and Scott [13] to be soluble in terms of an equivalence originally defined by Nerode [14]. However, even in this case, the added constraint that the complete set of language strings is not known, but only an *incomplete* set of positive and negative instances is available, is sufficient to turn it from a deductively decidable problem to one of *inductive* inference where the derived structure is only a *hypothesis constrained by the data* rather than a conclusion drawn from it.

The problem is further compounded if the underlying structure is known to be *probabilistic* so that only a distribution over language strings is available. In these circumstances the inference of the structure generating the behaviour becomes a statistical problem with no well-defined solution. For example, the string AAAABAAA could have been generated by a Bernoulli process (1-state stochastic automaton), or by a 5-state deterministic automaton, or by some other stochastic automaton. The decision between these hypotheses is an inductive one and requires assumptions not derived from the data. These create new problems, e.g. Gaines [15] shows that the assumption of deterministic causality for modelling data in fact from a probabilistic source does not lead to approximate models but in fact to *meaningless* ones that are just memories of the data.

There have been developed a variety of behaviour/structure inferencing systems based on heuristic techniques and applied to actual data [12]. Recently Gaines [1, 16, 17] gave a formulation and solution to the general system identification problem in terms of admissible *subspaces* of models ordered by *complexity* on the one hand, and by *approximation* to the observed behaviour on the other. He also specialized this with particular orderings appropriate to the problem of stochastic grammatical inference and demonstrated solutions to particular problems with these embodied in the computer program, ATOM. These specific results, and the related inferencing schemes of [18] and [19], require precise data. However, in real applications, e.g. animal ethology, the observations themselves may be vague or imprecise, and it is of interest to determine whether the inferencing techniques can be generalized to deal with data that are not known precisely.

In this paper the results of [1] are generalized to fuzzy systems in which only the degree of membership of a string to a language is known. The next section gives a synopsis of the results for precisely given data and the section following it generalizes this to fuzzy data.

2. General system identification

Feldman [20] pointed out that the selection of the 'best' structure for a particular behaviour was not well-defined even in the deterministic, complete sample, case. In general many structures may equally well fit the data and we need to define some preference relation over them in order to make the selection. This requirement for such a selection principle is a well-known one in the philosophy of science literature and William of Occam is generally credited with the principle of *choosing the simplest hypothesis.* For deterministic finite-state automata, for example, the Nerode equivalence leads to the minimal-state model of the data. For more complex structures, e.g. phrase-structure grammars, the preference ordering of simplicity or complexity is less obvious. Feldman notes that there are a number of possible orders on such grammars and, following [21], gives some desirable constraints on the semantics of 'complexity'.

When non-deterministic structures are considered also there is no longer a sense in which the best structure must exactly fit the data. It is reasonable only to suppose that the structure chosen will be a good *approximation* to the data. Horning paid particular attention to the requirements for measures of approximation in his work on grammatical inference [18], and Wharton [22] considers a variety of measures. The problem of behaviour-structure inference can now be seen as that of determining those models that are as good as possible in that no simpler, or equally simple, model is a better approximation to the data. Gaines [1] terms such models *admissable,* and formulates the identification problem in very general terms.

Suppose that we have two sets: B, of possible observed behaviours, and M, of possible models for behaviours; together with the pointed monoid, $(\mathrm{Ord}_M, \leqslant)$ of all pre-order relations on M, with one specified relation, $\leqslant$, singled out; and a mapping, $f: B \rightarrow \mathrm{Ord}_M$, from behaviours to orders on models. The quadruple, $(B, M, \leqslant, f)$ defines an *identification space*: the relation $\leqslant$ is one of model complexity such that if $m \leqslant n$, other things being equal, we should not prefer n to m; the mapping f is determined by further order relations of approximation that each behaviour induces on the set of models. We shall write $\leqslant_b$ for $f(b)$ so that if $m \leqslant_b n$ then m is not a worse approximation to the behaviour b than is n.

Now we are in a position to define a solution to the identification problem in terms of the product of the two pre-order relations, $\leqslant$ and $\leqslant_b$, which we shall define as $\leqslant_b^*$:

$$\forall m,n \in M,\ m \leqslant_b^* n \leftrightarrow m \leqslant n \quad \text{and} \quad m \leqslant_b n$$

i.e. $m \leqslant_b^* n$ if and only if m is neither more complex nor a worse approximation than n. The minimal elements in this order are all *admissible* solutions to the identification problem because they cannot be decreased in complexity without worsening approximation, and cannot be improved in approximation without increasing complexity. They form the admissible subspace determined by b, $M_b \subset M$, such that:

$$M_b \equiv \{m : \forall n \in M,\ n \leqslant_b^* n \leftrightarrow m \leqslant_b^* n\}$$

i.e. if any model is better than one in M_b then it is equivalent to it. Ralescu [23] has recently given a category-theoretic formulation of this condition.

2.1. Stochastic automata identification

Gaines [1] notes that the relations defining the identification problem are arbitrary, and recent results on computational complexity show that they are truly so in the sense that the Blum complexity classes can be arbitrarily ordered under quite strong semantic constraints [24]. However, in specializing the general result to specific problems intensional constraints may be applied suggesting particular orders, e.g. those which Zeigler considers for automata [25]. Stochastic automata have a natural complexity order in terms of their number of states. Regarded as grammars, they have an alternative natural order in terms of the number of possible transitions between states (which corresponds to the number of elements in the associated grammar).

There is also a range of possible measures of approximation between a given stochastic language and a possible stochastic automaton model. Maryanski [19] gives one in terms of a chi-square statistic. Gaines [1] derives a number of possible measures from the *subjective probability* eliciting schemes of Savage [26] and de Finetti [27]. These are part of a general family [28] whose underlying economic foundations Pearl has recently developed [29]. For the examples in this paper the logarithmic scoring rule of Savage will be used, but the others are also readily fuzzified.

All of the scoring rules give a measure of the discrepancy between the predictions of the automaton model and the events (i.e. words in language strings) that are actually observed. The logarithmic rule is that, if the word w_i actually occurs as event e_j and is predicted by the model to have a probability p_{ij} then the loss associated with the prediction is $-\log_2(p_{ij})$, and the total measure of approximation of the model to the behaviour is:

$$\mathrm{LE} = -\sum_j \sum_i \lambda_{ij} \log_2(p_{ij})$$

where $\lambda_{ij} = 1$ if event e_j is word w_i, and is 0 otherwise. The value of the loss at each prediction is a particularly interesting way of viewing the model's analysis of the data

$$C_j = -\sum_i \lambda_{ij} \log_2(p_{ij})$$

goes from 0 for a perfect prediction to infinity for a totally unexpected event (given probability 0)—we term it the *surprise* at the event.

Given the definitions of complexity and approximation, an enumerative inferencer may be designed which generates all models of lowest complexity, evaluates each in turn for approximation to the given behaviour, and outputs the model with best fit, then generates models of next higher complexity, and so on. The output of such an inferencer is the admissible sub-space of models in order of increasing complexity. The key design problem with such inferencers is to generate no models unnecessarily. Wharton [30] has given efficient generation schemes for a variety of grammars and complexity orderings.

ATOM, the stochastic structural inferencer described in [1], [16] and [17], uses an optimal enumeration procedure that generates models in terms of complexity (measured either by states or by transitions between them) in such a way that only possible models are generated and each is generated only once. ATOM allows a choice of measures of approximation including: maximum likelihood errors; Savage's logarithmic error defined above; Finetti's square error; Maryanski's chi-square; and total probability not accounted for. The measure used in this paper is LE as defined with an additional term to allow for the *derivational complexity* of the model:

$$\mathrm{TLE} = \mathrm{LE} + \log_2 (N_C)$$

where N_C is the total number of models enumerated up to and including the complexity class of the model concerned. If $1/N_C$ is regarded as the probability of the model itself, since $2^{-\mathrm{LE}}$ is the conditional probability of the model generating the behaviour, we have that the joint probability of model and behaviour is then $2^{-\mathrm{TLE}}$, so that the probability of the model given the behaviour is maximized when the TLE is minimized. Horning uses a similar measure but derives a probability for the model from a 'grammar-grammar' [18].

A typical result from ATOM is the derivation of a grammar for the sentences used by Feldman [31] and Evans [32]:

CAAAB, BBAAB, CAAB, BBAB, CAB, BBB, CB.

This is entered to ATOM as the sequence:

CAAAB/BBAAB/CAAB/BBAB/CAB/BBB/CB/ .

With the word / noted to be a "delimiter" such that it returns the model to its initial state. The resulting plot of minimal TLE against number of elements is shown in Fig. 1.

The derived 6-element grammar is:

$$\alpha \to B\beta \,|\, C\gamma \qquad \beta \to B\gamma \qquad \gamma \to A\gamma \,|\, B\delta \qquad \delta \to /$$

Note that even on this small sample the "best" grammar is well-defined and is an

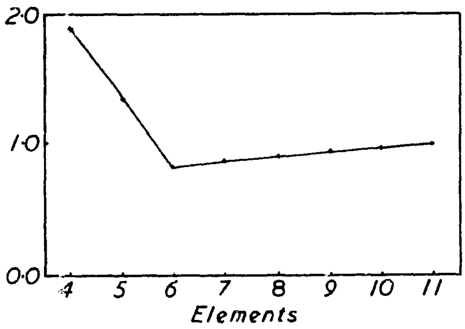

Fig. 1.

"inductive" one predicting sentences not yet observed, rather than the "deductive" 11-element one that produces just the actual sample.

3. Fuzzified general identification

The "fuzzification" [33, 34] of the formulation of general system identification given in Section 2 is fairly straightforward. Assume now that the behaviour is not observed precisely but is instead a "fuzzy restriction" [35] on the set of possible behaviours. More explicitly an observed behaviour is now a mapping, μ, from B to a "truth-set" V, $\mu : B \rightarrow V$. The truth-set is generally an ordered semiring [36], e.g. a lattice or the interval [0, 1] with max/min, add/multiply, or logical operations. Goguen [37] calls a mapping such as μ a V-set with B as carrier.

The mapping μ from B to V can clearly be extended in the usual way to M_b, the admissible subspaces, and hence to the models themselves. We may write:

$$\mu^*(m) = \bigvee_{B} (M_b(m) \wedge \mu(b))$$

where $M_b(m)$ is the characteristic function of M_b having as its value the maximum element in V if $m \in M_b$ and the minimum element otherwise, and $\vee$ and $\wedge$ are the semigroup operations on V, e.g. min and max respectively. The mapping $\mu^* : M \rightarrow V$ defines the fuzzy admissible subspace of models induced by the fuzzy behaviour $\mu : M \rightarrow V$.

This simple extension does not take into account the relative degrees of approximation of the same model to differing behaviours. In general it may not be possible to make such a comparison. However, if there is a uniform measure of approximation such that one can say that a model is a better approximation to one behaviour than to another, then the admissible subspace becomes a fuzzy restriction on the product of model and approximation spaces.

3.1. *Fuzzy stochastic automata identification*

ATOM, the specific stochastic automaton identifier described in Section 2.1. readily extends to the case where the data is a fuzzy language. Usually such a language will be generated by uncertainty about specific events so that it is convenient to generalize the λ_{ij} already defined to be the degree of membership of the event e_j to the word w_i. This generates a fuzzy restriction on the free semigroup of words, W^*, $\mu : W^* \rightarrow V$, such that if $x \in W^*$

$$\mu(x) = \bigwedge_{i,j} (\lambda_{ij} \wedge \eta_{ij})$$

where η_{ij} is max of V if the event e_j can be made equal to the corresponding word w_i in x, and min of V otherwise.

As described above the fuzzy restriction on data sentences (behaviours) may be used to derive a fuzzy restriction on models by considering the admissible set for each possible data sentence. This is best seen through an example—suppose the observed behaviour is: ABAB---BAB where the dashes represent uncertainty about the event which occurred; it might have been an A or B. This sequence might be represented by a series of degrees of membership:

A: 1	0	1	0	0.6	0.8	0.7	0	1	0
B: 0	1	0	1	1	1	1	1	0	1

a sequence of normalized fuzzy distributions [36]—complete uncertainty about A or B would result in them both being assigned degrees of membership 1.

This data generates the fuzzy language:

0.6: ABABAAABAB	0.6: ABABAABBAB
0.6: ABABABABAB	0.6: ABABABBBAB
0.7: ABABBAABAB	0.8: ABABBABBAB
0.7: ABABBBABAB	1 : ABABBBBBAB

ATOM may be envisaged as generating the admissible models for each sentence and giving them the associated degrees of membership. However, for this problem the TLE measures of approximation are comparable across the sentences and the procedure can also discard models whose degree of membership is exceeded by another of better approximation.

The results on this data are an admissible subspace of 3 probabilistic grammars:

$G_1(\mu^* = 0.6, \text{TLE} = 0.300)$

$\alpha \to A\beta\ (p = 1) \quad \beta \to B\alpha\ (p = 1)$

$G_2\ (\mu^* = 0.7, \text{TLE} = 0.693)$

$\alpha \to A\beta\ (p = 0.8) \mid B\beta\ (p = 0.2) \quad \beta \to B\alpha\ (p = 1)$

$G_3\ (\mu^* = 1, \text{TLE} = 0.817)$

$\alpha \to A\beta\ (p = 0.6) \mid B\beta\ (p = 0.4) \quad \beta \to B\alpha\ (p = 1)$

There are some interesting choices open in interpreting this data. If we precisify our data by only accepting a unity degree of membership then we have to adopt G_3 which ascribes the uncertainty in the data to a probabilistic source. If, however, we allow for imprecision in our observations by accepting a degree of membership 0.6 to admissible models we obtain G_1 as a deterministic model of our data. Note that the acceptance of G_1 takes advantage of the uncertainty in the data to simplify the model. It may be regarded as *precisiation* in Carnap's sense [38], i.e. removing vagueness in such a way as to promote the discovery of universal laws. In selecting G_1 we are saying: "What you actually observed as elements 5 through 7 was ABA, since this interpretation of your vague data enables me to tell you that there is a simple process of alternating A's and B's".

This procedure of generating probabilistic models of a fuzzy language is clearly very different from inferring fuzzy grammars [39]. We are not trying to *model* the imprecision in the data but are *using* it to precisiate the data to the best theoretical advantage. Such a process does seem very much akin to what is done in real-life research, and the fuzzy version of ATOM described brings us nearer an automated confirmation machine [40].

However, there is an alternative view that is also of interest. Consider the event by event predictions of the 3 models. They all agree that every second event is quite definitely B. However, they assign different probabilities to words at the intervening events, e.g. to the third event:

G_1: $\mu^*=0.6$ $p_A=1$ $p_B=0$
G_2: $\mu^*=0.7$ $p_A=0.8$ $p_B=0.2$
G_3: $\mu^*=1$ $p_A=0.6$ $p_B=0.4$

Following Zadeh [41] we may describe this event as giving a *possibility vector* to:

A of (1/0.6, 0.8/0.7, 0.6/1) and to
B of (0/0.6, 0.2/0.7, 0.4/1)

which might have linguistic approximations: A is *very likely* and B is *rather unlikely*. Thus the fuzzy ATOM procedure may be regarded as deriving *possibilistic models* of fuzzy sequential systems.

Thus the modelling schema described in this paper is itself neutral in that the models derived may be further restricted by criteria of precisiation, or may be interpreted without such restriction in possibilistic, and hence fuzzy linguistic, terms. In terms of the data itself no resolution between these views is possible—they represent legitimate, related, but essentially differing, *views of the world.*

4. Conclusions

This note is intended as a further step towards a unified theory of uncertainty for general systems. It demonstrates that Zadeh's theory of possibilistic systems, combining probability and fuzziness, may be developed operationally through computational algorithms for fuzzy sequential system identification. The modelling technique also throws light on the role of precisiation in scientific inductive inference.

References

[1] B.R. Gaines, System identification, approximation and complexity, Internat. J. General Systems 3 (1977) 145-174.

[2] L.A. Zadeh, On the identification problem, IRE Trans. Circuit Theory 3 (1956) 277-281.

[3] P. Eykhoff, System Identification (Wiley, London, 1974).

[4] R. Dawkins and M. Dawkins, Some descriptive and explanatory stochastic models of decision-making, in: D.J. McFarland (ed), Motivational Control Systems Analysis (Academic Press, 1974) 119-168.

[5] D.M. Vowles, Neuroethology, evolution and grammar, in: L.R. Aronson et al. (eds), Development and Evolution of Behaviour (Freeman, San Francisco, 1970) 194-215.

[6] P.J.B. Slater, Describing sequences of behavior, in: P.P.G. Bateson and P.H. Klopfer, Perspectives in Ethology 1 (Plenum Press, New York, 1973) 131–153.

[7] B.R. Gaines, Linear and nonlinear models of the human controller, Internat. J. Man-Machine Studies 1 (1969) 333–360.

[8] N. Chomsky, Three models for the description of language, IRE Trans. Information Theory 2 (1956) 113–124.

[9] R. Solomonoff, A new method for discovering the grammar of phrase-structure languages. Proc. Int. Conf. Information Processing, UNESCO Paris, Butterworths (1959) 285–290.

[10] R. Solomonoff, A formal theory of inductive inference, Information and Control 7 (1964) 1–22 and 224–254.

[11] E.F. Moore, Gedanken experiments on sequential machines, in: Automata Studies, Annals of Mathematical Studies 34 (Princeton University Press, New Jersey, 1956) 129–153.

[12] K.S. Fu and T.L. Booth, Grammatical inference: introduction and survey, IEEE Trans. Systems, Man and Cybernetics 5 (1975) 95–111 and 409–423.

[13] M.O. Rabin and D. Scott, Finite automata and their decision problems, IBM J. Research and Development 3 (1959) 114–125.

[14] A. Nerode, Linear automaton transformations, Proc. American Mathematical Society 9 (1958) 541–544.

[15] B.R. Gaines, On the complexity of causal models, IEEE Trans. Systems, Man and Cybernetics, 6 (1976) 56–59.

[16] B.R. Gaines, Approximate identification of automata, Electronics Letters 11 (1975) 444–445.

[17] B.R. Gaines, Behaviour/structure transformations under uncertainty, Internat. J. Man-Machine Studies 8 (1976) 337–365.

[18] J.J. Horning, A study of grammatical inference, Ph.D. thesis, Stanford University (1969) (University Microfilms 70-10, 465).

[19] F.J. Maryanski, Inference of probabilistic grammars, Ph.D. thesis, University of Connecticut (1974) (University Microfilms 75-10, 645).

[20] J.A. Feldman, Some decidability results on grammatical inference and complexity, Information and Control 20 (1972) 244–262.

[21] M. Blum, A machine-independent theory of the complexity of recursive functions, J. Association for Computing Mahinery, 14 (1967) 322–336.

[22] R.M. Wharton, Approximate language identification, Information and Control 26 (1974) 236–255.

[23] D. Ralescu, Approximat, models for system identification (Faculté de Science Economiques et de Gestion, Université de Dijon, France, 1977).

[24] R. Moll, An operator embedding theorem for complexity classes of recursive functions, Theoret. Comput. Sci. 1 (1976) 193–198.

[25] B.P. Zeigler, Simulation based structural complexity of models, Internat. J. General Systems 2 (1976) 217–223.

[23] B.P. Savage, Elicitation of personal probabilities and expectations, J. Am. Statist. Assoc. 66 (1971) 783–801.

[27] B. De Finetti, Probability, Induction and Statistics (Wiley, London) 1972.

[28] J. Aczel and J. Pfanzagl, Remarks on the measurement of subjective probability and information, Metrika 2 (19) 91–105.

[29] J. Pearl, An economic basis for certain methods of evaluating probabilistic forecasts, Internat. J. Man-Machine Studies 10 (1978) 175–183.

[30] R.M. Wharton, Grammar enumeration and inference, Information and Control, 33 (1977) 253–272.

[31] J.A. Feldman, J. Gips, J.J. Horning and S. Reder, Grammatical inference and complexity, Artificial Intelligence Memo (Computer Science Dept. Stanford University, 1969).

[32] T.G. Evans, Grammatical inference in pattern analysis, in: J.T. Tou (ed.), Software Engineering, Vol. 2 (Academic Press, New York, 1971) 183–202.

[33] B.R. Gaines, Foundations of fuzzy reasoning, Internat. J. Man-Machine Studies 8 (1976) 623–668.

[34] B.R. Gaines, and L.J. Kohout, The fuzzy decade: a bibliography of fuzzy systems and closely related topics, Internat. J. Man-Machine Studies 9 (1977) 1–68.

[35] L.A. Zadeh, Calculus of fuzzy restrictions, Memo ERL-M502, Electronics Research Laboratory, College of Engineering, University of California, Berkeley (Feb. 1975).

[36] B.R. Gaines and L.J. Kohout, The logic of automata, Internat. J. General systems 2 (1976) 191–208.

[37] J.A. Goguen, Concept representation in natural and artificial languages: axioms, extensions and applications for fuzzy sets, Internat. J. Man-Machine Studies 6 (1974) 513–561.

[38] R. Carnap, Logical Foundations of Probability (University of Chicago Press, 1950).
[39] N. Honda and M. Nasu, Recognition of fuzzy languages, in: L.A. Zadeh, K.S. Fu, K. Tanaka and M. Shimura (eds.), Fuzzy Sets and Their Applications to Cognitive and Decision Processes (Academic Press, New York, 1975) 279–299.
[40] E. Erwin, The confirmation machine, in: R.C. Buck and R.S. Cohen (eds.), PSA 1970: In memory of Rudolf Carnap (Reidel, Dordrecht, Holland, 1971) 306–321.
[41] L.A. Zadeh, Fuzzy sets as a basis for a theory of possibility, Fuzzy sets and systems 1 (1978) 3–28.