Presented at 5th International Congress on Logic, Methodology and Philosophy of Science: London, Ontario, Canada, Aug. 27, 1975

THE ADEQUACY OF COMPLEXITY MODELS OF INDUCTION

R. Solomonoff, Zator Co., Camb., Mass., USA, 02138

We will discuss theorems showing that for a broad class of stochastic sources, Willis' model gives error rates close to minimum obtainable. Used as the basis of a betting system it again gives close to maximum yield. Cover's model, using Chaitin complexity, converges to the same optima but we are uncertain if the rate of convergence is as rapid.

Inductive inference can be defined to be the extrapolation of a binary sequence containing all the data to be used in the induction. Almost all, if not all, activities considered to be induction can be put into this form. The Algorithmic complexity of a sequence is related to the shortest programs for a reference universal Turing machine producing that sequence as output.

Earliest application of algorithmic complexity was for inductive inference[1] and it was shown that this complexity was relatively insensitive to choice of reference machine. Kolmogorov[2] and Chaitin[3] independently used somewhat different definitions of algorithmic complexity to define randomness of finite strings. Willis[4] later developed the original induction system in more rigorous form and gave many theorems that greatly clarified its operation. Cover[5], apparently independently of Willis, used Chaitin's complexity for induction and employed a simple betting system to measure its efficiency.

A model equivalent to Willis' uses as reference a universal Turing machine having the "sequential property"- i.e. if input string a gives A as output, then any input of the form $a^\frown b$ must have output of form $A^\frown B$. Willis assigns to string a, the probability

$$P_w(a) = \lim_{\ell \to \infty} N_\ell(a)/N_\ell \qquad (1)$$

$N_\ell(a)$ is the number of input strings of length ℓ that result in output strings of the form $a^\frown b$, where b may be any finite (including the null) string. N_ℓ is the number of input strings of length ℓ that give outputs that are at least as long as a is.

Cover's model is based on Chaitin's complexity using a universal prefix machine as reference, assigning to a the probability

$$P_c = A\sum_i 2^{-C_i} \qquad (2)$$

C_i is the Chaitin complexity of the ith possible continuation of a, and A is a constant that assures that the total probability of all strings of a given length is unity.

While both of these definitions of probability are not effectively computable, there exist sequences of computable probability estimates that converge to equations (1) and (2).

Suppose we have a binary sequence that has been created by a stochastic generator. Through Bayes' theorem each of the two systems can be used to give the conditional probability of each bit in the sequence, given the antecedent subsequence.

First, we will use each system as the basis of a betting scheme at even odds. Suppose we start with a fortune of unity and bet a fraction p_i^1 that the next bit will be 1 and a fraction $p_i^0 = 1-p_i^1$ that the next bit will be 0. At the end of n bets, our total fortune will be exactly $2^n P_i(x(n))$. Here x(n) is the binary sequence that has occurred, and $P_i(x(n))$ is the probability that our system has assigned to that sequence.

The log of Our fortune has a maximum expected value if $P_t(x(n))$ is the probability that would be assigned by the stochastic generator. If P_i is a less perfect probability value the fortune will be somewhat less - depending on how good P_i is. One goodness criterion of a system which we'll call b_i, is the log of the ratio of its betting yield to the maximum possible.

$$b_i = \log_2 (P_i(x(n))/P(x(n)) \qquad (3)$$

Another criterion considers the differences in conditional probability obtained by P and by P_i. Let δ_n^i be the conditional probability by P_i of the nth bit of x(n), given the previous n-1 bits, i.e.

$$\delta_n^i \equiv P_i(x(n))/P_i(x(n-1))$$

Then the expected value of $\sum_{j=1}^{n} (\delta_j^i - \delta_j)^2$ is a measure of the error in P_i. It can be shown that

$$E\left(\sum_{j=1}^{n}(\delta_j^i-\delta_j)^2\right) \equiv \sum_{k=1}^{2^n}\left(P({}^kx(n))\sum_{j=1}^{n}({}^k\delta_j^i-{}^k\delta_j)^2\right) \leq b_i \ln\sqrt{2}$$

E is the expected value with respect to P. ${}^kx(n)$ is the kth sequence of length n (there are just 2^n of them). ${}^k\delta_j^i$ and ${}^k\delta_j$ are the conditional probabilities of the jth bits of ${}^kx(n)$ for P_i and P, respectively. b_i is the same as in equation (3).

If the underlying stochastic generator is describable in a finite number of bits, d, then for $P_t = P_w$ (Willis' system), $b_w = d$.

If the generator is finitely describable, except for k differentiable parameters, (a differentiable parameter has an infinitely long description), then if the functional form is known and the parameters are known to within a certain error, b_t will be of the form c log n, c being a constant. For such a generator, Willis' system will give $b = c \log n + d$. c is the same constant as before, and d is the number of bits in the description of the functional form of the stochastic generator.

If the generator is ergodic, then Cover has shown that for his system,

$$\lim_{n \to \infty} \frac{b_c}{n} = 0$$

In Willis' system $\frac{b_w}{n}$ also approaches zero but at a known rate, probably as fast as is theoretically possible.

We don't yet have bounds on how rapidly $\frac{b_c}{n}$ approaches zero - conceivably, it could do so very slowly. Preliminary analysis, however, suggests that the approach might be as rapid as that of Willis' method.

The systems described seem quite adequate for prediction and cast much light on some classic problems in inductive inference theory.

One possible approach to defining randomness of a finite sequence is that all future continuations of the sequence are about equally likely. The forgoing systems make it possible to put such a definition into an exact form and analyse its properties.

The problem of geometric probability is best understood by converting the data to digital form (using most any analog to digital conversion method) and analysing the data as a binary sequence. Various measure transformations have a simple interpretation in this light, and their significance (or lack of significance) in modifying probability values can be readily evaluated.

Goodman's paradoxes involving various linguistic transformation of inferential data, are also easy to analyse from this point of view.

The most important open problems in induction theory are in the practical realization of systems such as have been described. We want to obtain the most accuracy in induction for a given amount of computing. This involves optimizing a function of algorithmic and computational complexities - a kind of problem that occurs in information retrieval and in some models of organic evolution.

References

1. Solomonoff, R.J. "A Formal Theory of Inductive Inference." Information and Control, March 1964, pp. 1-22, June 1964, pp. 224-254.

2. Kolmogorov, A.N. "Logical Basis for Information Theory and Probability Theory." IEEE Transactions on Information Theory IT-14, pp. 662-664, 1968.

   "Three Approaches to the Quantitative Definition of Information," Information Transmission, Vol. I, pp. 3-11, 1965.

3. Chaitin, G.J. "A Theory of Program Size Formally Identical to Information Theory." To appear in Journal of the Association of Computing Machinery, July, 1975 pp. 329-340.

4. Willis, D.G. "Computational Complexity and Probability Constructions," Journal of the Assoc. of Computing Machinery, April 1970, pp. 241-259.

5. Cover, T.M. "Universal Gambling Schemes and the Complexity Measures of Kolmogorov and Chaitin," Report No. 12, Statistics Dept., Stanford University, Stanford, Calif., 1974. Submitted to Annals of Statistics.