Learning Sequential Patterns for Probabilistic Inductive Prediction

Keith C. C. Chan, Andrew K. C. Wong, *Member, IEEE*, and David K. Y. Chiu, *Member, IEEE*

***Abstract*—Suppose we are given a sequence of events that are generated probabilistically in the sense that the attributes of one event are dependent, to a certain extent, on those observed before it. This paper presents an inductive method that is capable of detecting the inherent patterns in such a sequence and to make predictions about the attributes of future events. Unlike previous AI-based prediction methods, the proposed method is particularly effective in discovering knowledge in ordered event sequences even if noisy data are being dealt with. The method can be divided into three phases: (i) detection of underlying patterns in an ordered event sequence; (ii) construction of sequence-generation rules based on the detected patterns; and (iii) use of these rules to predict the attributes of future events. The method has been implemented in a program called OBSERVER-II, which has been tested with both simulated and real-life data. Experimental results indicate that it is capable of discovering underlying patterns and explaining the behaviour of certain sequence-generation processes that are not obvious or easily understood. The performance of OBSERVER-II has been compared with that of existing AI-based prediction systems, and it is found to be able to successfully solve prediction problems programs such as SPARC have failed on.**

I. Introduction

MUCH research work in inductive learning addresses the problem of *classification*. Given a collection of objects (events, observations, situations, processes, etc.) described in terms of one or more *attributes* and preclassified into a set of known classes, the classification problem is to find a set of characteristic descriptions for these classes. Or, equivalently, a procedure for identifying an object as either belonging to or not belonging to a particular class. If each class of objects is considered as exemplifying a certain concept, then, a system that is capable of sorting those objects belonging to a class from those that do not, can be considered to have acquired the concept associated with the class [30].

Based on the discovery that there is a discrepancy between an individual's reading and speaking vocabularies, Simon and Kotovsky [30], however, observed that there is no necessary relationship between the ability of a learning system to identify an object as belonging to a concept, and its ability to produce examples of that concept. They noted that the acquisition of certain kinds of concepts – such as those in the form of serial patterns – can only be measured by a system's ability to produce an object satisfying the concept, rather than its ability to determine if an object exemplifies it. For instance, a system is considered to have acquired the concept 'simple alternation of a and b' in 'abababab__' only if it can extrapolate the letter series by producing the succeeding characters (i.e., 'ba').

Since a learning system that is able to acquire the concept embedded in a sequence of objects is also able to predict the characteristics of future objects based on the acquired concept, such a task may be referred to as prediction. More formally, the prediction task can be stated in the following way. Suppose that we are given an ordered sequence of objects (observations, events, situations, phenomena, etc.) described by one or more attributes. Suppose also that these objects are generated by a certain process in such a way that the attributes of one object are dependent on those of the preceding objects. The prediction task is, therefore, to find a set of characteristic descriptions of the sequence so that, based on these descriptions, the characteristics of future objects can be predicted.

As an illustration of the prediction task, let us suppose that we are given a snapshot of an ongoing process which has generated a sequence of locomotives shown in Fig. 1.

Suppose that each locomotive is characterized by four attributes: NUMBER OF WHEELS with values {Two, Three, Four}, LENGTH OF FUNNEL with values {Long, Short, Medium}, NUMBER OF STRIPES with values {One, Two, Three}, and NUMBER OF WINDOWS with values {One, Two, Three}. If the attributes of each locomotive in the sequence are dependent, to a certain degree, on those preceding it, then the prediction problem is to find the rules that governs the generation of the sequence and to use these rules to predict the characteristics of future locomotives.

Prediction problems can be categorized into two different types: the *deterministic prediction* (*DP*) problem and the *probabilistic prediction* (*PP*) problem. If all the attributes of an object in a sequence can be predicted with complete certainty based on the preceding ones, then the prediction problem is deterministic. However, if due to the inherent random nature of the process that generates the data or to missing, inconsistent, or inaccurate values, an attribute of an object can only be predicted with some degree of certainty based on those preceding it, then the prediction problem is probabilistic. For example, predicting future letters in a sequence such as 'abababababa__' is deterministic, since they are completely determined by the existing ones; predicting weather conditions based on past observations is probabilistic,

Manuscript received May 25, 1990; revised December 1, 1992 and December 10, 1993.

K. C. C. Chan is with the Dept. of Electrical & Computer Engineering, Ryerson Polytechnic University, 350 Victoria Street, Toronto, Ontario, Canada M5B 2K3.

A. K. C. Wong is with the PAMI Laboratory, Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

D. K. Y. Chiu is with the Department of Computing and Information Science, University of Guelph, Guelph, Ontario, Canada, N1G 2W1.

IEEE Log Number 9403049.

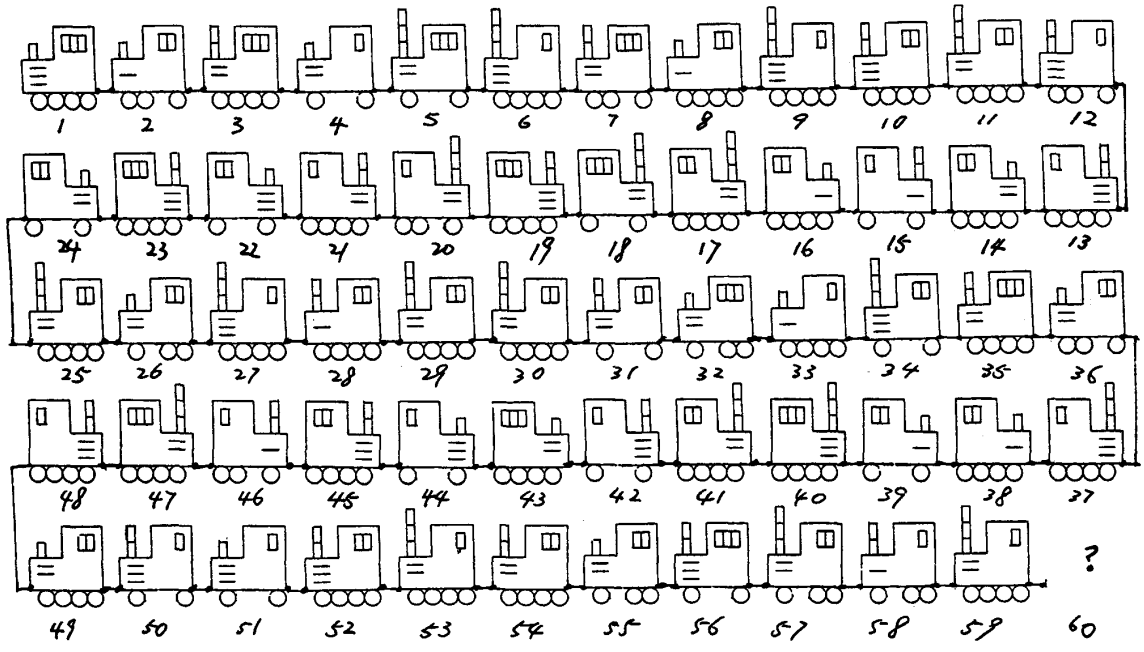

Fig. 1. A prediction problem involving a sequence of locomotives.

since weather forecasts cannot usually be made with complete certainty.

Thus the PP problem is, in general, more difficult to solve than the DP problem, due to the need to consider probabilistic variations. It is, however, a more important problem to be tackled. This is because the ability to make accurate predictions in the presence of uncertainty is very important in daily life and is crucial in a wide range of disciplines such as business, economics, sociology, medicine, science and engineering. An efficient solution to the PP problem would be useful for tasks such as predicting weather conditions, earthquakes, or volcanic activities; or in the business world for predicting and monitoring the consumer index, the behaviour of stock markets, etc.

The need for an efficient solution to the PP problem is further evidenced by the construction of expert systems in such areas of science and engineering [24], [31], meteorology and climatology [1], [11], [21], [28], [32], [38], and business and finance [7], [18]. Such systems help decision-makers to forecast changes in the future based on knowledge of the past. They are built by knowledge engineers laboriously extracting domain knowledge from experts through interviews, and their construction is, therefore, difficult and time-consuming. Furthermore, since human judgement is often found to be inaccurate and inconsistent in the presence of uncertainty, the extracted knowledge is usually also of questionable quality [36].

To overcome these problems, this paper proposes an efficient solution strategy to the PP problem based on an AI-based inductive learning method This method differs from existing ones in the following ways: (i) it is data-driven instead of model-driven; (ii) it does not require the characteristics of the objects in a sequence to be strictly a function of that of those preceding them, and (iii) it is able to effectively handle noisy data. Based on a newly developed technique for probabilistic inference [2], [3], [4], [6], the proposed method can efficiently deal with the PP problem by uncovering the hidden relationships between the attributes of objects located at different positions in a sequence. It is also able to (i) quantitatively estimate the reliability of these relationships, (ii) combine the evidence of all indicators in order to provide the best prediction, and (iii) estimate the confidence in such prediction. These capabilities of the proposed prediction method enables objective prediction knowledge to be acquired rapidly, even when some data values are missing or incorrect.

The proposed method has been implemented in a system known as OBSERVER-II. The OBSERVER-II has been tested with both simulated and real-world data and its performance is found to be better, in terms of (i) computational efficiency and (ii) predictive accuracy, than some well-known AI-based learning systems that deal with the prediction problems.

II. Uncertainty Handling in Existing Prediction Systems

The prediction problem was first considered by some psychologists, with the goal of simulating human cognitive processes. For instance, Simon and Kotovsky proposed a model

of human behaviour in solving letter-series extrapolation problems [19], [30]. Based on their model, a computer program known as the *Concept Former* was constructed. This program is able to discover periodic patterns implicit in some letter sequences by searching for letter pairs that are related in the sense that they are the 'same' or are 'next' to each other in the alphabet. If such a relationship is found to repeat or to be interrupted at regular intervals, the program establishes a period boundary. Once such basic periodicity within a sequence is determined, the details of the patterns are acquired by detecting if this relationship holds between successive letters within a period or between letters in corresponding positions of successive periods. These acquired patterns are then used to extrapolate the letter sequence.

A number of other programs have also been written to handle letter-series extrapolation problems. As with Concept Former, these programs acquires a description of a letter series by attempting to determine whether or not simple relationships, such as 'same', 'next', 'next after next', or 'predecessor', exist between letters [12], [26] [27], [35]. Even though these programs were developed primarily for handling prediction problems involving letter series, they can also be used for number-sequence extrapolation. These programs can be modified slightly so that, other than the relations of 'same', 'next' and 'previous' between numbers, arithmetical relationships such as 'addition', 'subtraction', 'exponentiation', etc. are also considered [15], [17], [20]. Fredkin, for example, used the idea of 'successive difference' (SD) to determine if a number series is generated by a polynomial of degree n [15]. But his program is not able to handle polynomial approximations. Pivar and Finkelstein modified it later to deal with simple polynomials with exceptions (e.g. If $n = 10$, then $f(n) = 33$, else $f(n) = 3n$, etc.) [23].

If a number sequence fits neither a simple polynomial equation nor one with exceptions, then neither Fredkin nor Pivar and Finkelstein's program will work. In such case, Persson's approach [26] can be adopted. By applying Newton's forward-difference formula, this approach is able to compute the coefficients and the degree of a polynomial approximation of a number sequence.

However, none of these programs that were originally developed for letter- or number-series extrapolation can be used to deal with prediction problems involving noisy data. Furthermore, they are limited to handling DP problems in which objects in a sequence are characterized by only one attribute – a letter or a number. Their ability to solve real-world problems is further restricted by the fact that arithmetic relations or relations such as 'next', 'previous', or 'next after next' can only be meaningfully defined between numbers or letters; they are not able to uncover hidden patterns in object sequences characterized by symbolic data measured according to the nominal scale

To deal with DP problems other than number- or letter-series extrapolation, AI-based inductive learning programs such as the THOTH program [33] can be used. THOTH is capable of determining maximal conjunctive generalizations between pairs of objects in a sequence so as to construct a set of generalized descriptions of the serial patterns underlying it. Even though THOTH is capable of handling prediction problems involving objects described by multiple attributes, it has the disadvantage that it is relatively slow. Furthermore, it can only handle prediction problems in which all the characteristics of the next object in a sequence can be determined exactly, and with complete certainty, based on the previous ones. To deal with uncertainty in prediction, the SPARC program was developed [10], [22].

SPARC is capable of handling a certain type of PP problem in which a sequence of objects is assumed to be described by two different types of attributes: (i) those whose values cannot be determined with complete certainty, and (ii) those whose values can be determined completely based solely on the attribute values of the previous objects. For attributes whose values cannot be deterministically predicted, SPARC either assumes them to be of the second type and finds an over-fitted model, or refuses to make any predictions. For attributes whose values can be deterministically predicted, SPARC employs a model-directed approach to guide the search for suitable prediction rules in a predetermined search space which consists of all possible rule models.

Depending on the application domains, the number of rule models that have to be considered by SPARC during this searching process may be very large. For example, in an attempt to discover the rules that govern the generation of a sequence of playing cards in a card game, as many as 10^{137} possible rules have to be considered [10]. Since a breadth-first search of such a huge rule space would be impossible, the size of the search space has to be restricted. The following assumptions are therefore made by SPARC: (i) there are three different types of sequence-generating rules: periodic, decomposition, and disjunctive normal form rules, (ii) one sequence can be transformed into another by segmenting, splitting, and blocking, so as to determine if its generation process fits these models, (iii) the rule consists of a certain number of conjunctive terms on the left-hand side, and (iv) substantial amounts of domain-specific knowledge are available to guide the searching process.

In problem domains where domain-specific knowledge is unavailable and the assumptions about the possible rule models and the different types of sequence transformations cannot be validly made, SPARC would, therefore, be unable to discover the sequence-generation rules. And it is, for this reason that it cannot make predictions about attribute values of an object that are not completely determined by the values of those preceding it. A further limitation of SPARC is its use of a nearly exhaustive search strategy in finding suitable models which makes the learning process it employs rather slow. Also, if the objects in a sequence are generated probabilistically, as in most PP problems, either SPARC refuses to make predictions due to the lack of a suitable model in the search space, or else the attributes are treated as completely deterministic. This not only leads to overfitting, but also to an exploded search space consisting of many plausible rule models that would render the model-driven learning approach employed by SPARC unfeasible. For these reasons, SPARC cannot be practically used in application domains where complete certainty is unattainable.

Other than the cognitive-model based approaches for solving the letter- or number-series extrapolation problems; the THOTH program for solving DP problems involving object sequences described by one or more attributes; and the SPARC program for solving a special type of PP problems involving sequences describable by certain kinds of rule models, not much work on either the DP or the PP problems has been reported in the AI literature. There have been efforts to find suitable methods for modeling and reasoning about dynamically changing systems so as to better predict their behaviour [8], [14]. But, instead of acquiring general descriptions inductively from partial knowledge of a certain event, these methods have mainly been concerned with deductive inference. In other words, the prediction problems that these systems are trying to tackle are not inductive in nature, and are hence different from the type of DP or PP problems that we are concerned with.

The works of some statisticians to analyze noisy time series data is closer in objective to the work described in this paper. Unfortunately, time series analysis techniques were mainly developed for prediction problems involving quantitative variables. They cannot be modified to deal with PP problems where (i) the knowledge or concept to be acquired must be expressed in symbolic form, and (ii) predictions of qualitative variables have to be made based on symbolic representations of object sequences.

In summary, existing prediction systems can only handle DP problems. Even though SPARC can be considered an exception, it is unable to deal with problems involving data that are generated by a probabilistic process or are characterized by missing, erroneous, or inconsistent values. To efficiently handle PP problems in general, a simple yet effective learning method that is considerably different from those described above, is proposed. This method is based on a newly developed probabilistic inference technique [2], [3], [4], [6], and is able to make predictions about a future object whose attributes are dependent on those of the observed ones only to a certain degree.

III. Description of the Prediction Problem

The probabilistic prediction problem can be described more formally as follows: suppose that there is an ordered sequence S of M objects, $obj_1, \ldots, obj_p, \ldots, obj_M$, where obj_p is located at position p in S. Suppose also that each object in the sequence is described by n distinct attributes, $Attr_{1p}, \ldots, Attr_jp, \ldots, Attr_{np}$, and that in any instantiation of the object description, an attribute $Attr_{jp}$ takes on a specific value, $val_{jp} \in domain(Attr_{jp}) = \{v_{j_k} \mid k = 1, \ldots, J\}$, which may be numerical or symbolic, or both.

In the presence of uncertainty, it is possible for some of the attribute values that characterize the objects in S to be missing or erroneous. It is also possible for the process that generates S to have some inherently random element. In either case, the attributes of the objects in the sequence can be considered as dependent probabilistically on those preceding it. Given such a sequence of objects, the PP problem is to find a set of prediction rules that describes how S is generated and that can be employed to predict the characteristics of a future object.

IV. An Inductive Method for the Learning of Prediction Rules

Having described the PP problem, we will now propose an inductive learning method to solve it. This method consists of three phases: (i) detection of patterns inherent in a sequence of objects, (ii) construction of prediction rules based on the detected patterns, and (iii) use of these rules to predict the characteristics of future objects.

A. Detection of Sequential Patterns in Noisy Training Data

For accurate predictions, it is important to know how the attribute values of the objects in a sequence are dependent on the preceding ones. If the ith attribute of an object that takes on v_{i_l} is always preceded at τ positions (or time units) earlier by an object whose jth attribute takes on the value v_{j_k}, one can conclude that v_{i_l} is dependent on v_{j_k} with a position (or time) lag of τ. However, if it is observed that v_{j_k} is never followed by v_{i_l} at τ positions later, we can also conclude that v_{i_l} is dependent (in a negative sense) on v_{j_k} with a position (or time) lag of τ. That is, whenever an object is observed to have the characteristic v_{j_k}, the object that is located at τ positions later in the sequence will not possess the attribute value v_{i_l}. In either case, the jth attribute can be considered as providing useful information for the prediction of future objects.

In the presence of noise in the data, the identification of such attributes is not easy. This is because the correlation between two attributes is rarely perfect. To make accurate predictions in uncertain environments and to avoid overfitting, the presence of counter-examples should be tolerated to a certain extent. Instead of requiring the correlation between the attributes of objects located at different positions in a sequence to be perfect, an attribute, say $Attr_{jp}$, should therefore be regarded helpful in determining the attributes of future objects as long as those attributes are dependent on it probabilistically.

To decide if the ith attribute of an object in a sequence is dependent on the jth attribute $Attr_{jp}$ of the object at τ positions earlier, the chi-square test can be employed. A two-dimensional contingency table (Table I) of I rows and J columns (I and J being the total number of values taken on by the ith and the jth attributes, respectively) can be constructed.

Let o_{lk} be the total number of objects in S whose ith attribute, $Attr_{i(p+\tau)}$, takes on the value v_{i_l} and are preceded at τ positions earlier by objects that have the characteristic v_{j_k}. Let e_{lk} be the expected number of such objects. Under the assumption that $Attr_{i(p+\tau)}$ and $Attr_{jp}$ are independent, $e_{lk} = \sum_{u=1}^{J} o_{lu} \sum_{u=1}^{I} o_{uk}/M'$, where $M' = \sum_{l,k} o_{lk}$ is less than or equal to M (the total number of objects in the sequence S) due to the possibility of there being missing values in the data. A chi-square statistic can then be defined as:

$$X^2 = \sum_{l=1}^{I} \sum_{k=1}^{J} \frac{(o_{lk} - e_{lk})^2}{e_{lk}} = \sum_{l=1}^{I} \sum_{k=1}^{J} \frac{o_{lk}^2}{e_{lk}} - M'. \quad (1)$$

Whether or not the difference between what is observed and what is expected could have arisen by chance can be determined by comparing the observed chi-square statistic X^2 with the critical chi-square ${\chi^2}_{d,\alpha}$, where $d = (I-1)(J-1)$ is

TABLE I
A TWO-DIMENSIONAL CONTINGENCY TABLE WITH I ROWS AND J COLUMNS

		$Attr_{jp}$						
		v_{j_1}	v_{j_2}	$\cdots$	v_{j_k}	$\cdots$	v_{j_J}	Totals
	v_{i_1}	o_{11} (e_{11})	o_{12} (e_{12})	$\cdots$	o_{1k} (e_{1k})	$\cdots$	o_{1J} (e_{1J})	o_{1+}
	v_{i_2}	o_{21} (e_{21})	o_{22} (e_{22})	$\cdots$	o_{2k} (e_{2k})	$\cdots$	o_{2J} (e_{2J})	o_{2+}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$Attr_{i(p+\tau)}$	v_{i_l}	o_{l1} (e_{l1})	o_{l2} (e_{l2})	$\cdots$	o_{lk} (e_{lk})	$\cdots$	o_{lJ} (e_{lJ})	o_{l+}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	v_{i_I}	o_{I1} (e_{I1})	o_{I2} (e_{I2})	$\cdots$	o_{Ik} (e_{Ik})	$\cdots$	o_{IJ} (e_{IJ})	o_{I+}
	Totals	o_{+1}	o_{+2}	$\cdots$	o_{+k}	$\cdots$	o_{+J}	M'

TABLE II
A 3 × 3 CONTINGENCY TABLE FOR NUMBER OF WINDOWS AND NUMBER OF WHEELS

	o_{lk} (e_{lk})	NUMBER OF WHEELS Two	Three	Four	TOTAL
NUMBER OF WINDOWS	One	3 (5.31)	7 (4.17)	12 (12.52)	22
	Two	4 (5.79)	1 (4.55)	19 (13.66)	24
	Three	7 (2.90)	3 (2.28)	2 (6.83)	12
	TOTAL	14	11	33	58

the degrees of freedom and α, usually taken to be 0.05 or 0.01, is the significance level ($(1-\alpha)\%$ is the confidence level). If X^2 is greater than the critical value, there is enough evidence to conclude that $Attr_{i(p+\tau)}$ is dependent on $Attr_{jp}$; otherwise, if X^2 is less than $\chi^2{}_{d,\alpha}$, one cannot conclude this [29].

As an illustration of the chi-square test, let us consider the problem of predicting the characteristics of locomotives (Fig. 1). To determine if the attribute NUMBER OF WHEELS of a locomotive is important for predicting the attribute NUMBER OF WINDOWS of the next locomotive, a contingency table with three rows (since a locomotive can only have one, two, or three windows) and three columns (since a locomotive can only have two, three, or four wheels) can be constructed (Table II). Based on (2) above, the value of the chi-square statistic, X^2, is 17.82.

Since $X^2 = 17.82$ is greater than the critical chi-square values $\chi^2{}_{4,0.05} = 9.49$ and $\chi^2{}_{4,0.01} = 13.28$, the chi-square test is significant at both the 95% and 99% levels. This suggests that the attribute NUMBER OF WINDOWS of a locomotive is dependent on that of the attribute NUMBER OF WHEELS of the previous one, and we can, therefore, conclude that the latter is important in determining the former.

It should be noted, however, that even though a significant overall chi-square test allows us to conclude that an attribute, say $Attr_{i(p+\tau)}$, is dependent on another one, say $Attr_{jp}$, it provides no information as to how the observed values of the ith attribute of an object in a sequence is dependent on that of the jth attribute of the object at τ positions earlier. For instance, even though we know, by a significant chi-square test, that the number of wheels of a locomotive is important in determining the number of windows of the next locomotive in the sequence, we are unable to draw any conclusion as to whether a given locomotive with four wheels should be followed by one with three windows or not. In view of the importance of such information – especially when both the ith and the jth attribute take on a large number of different values – we propose a method to evaluate if a specific value, v_{i_l}, of the ith attribute of an object is statistically dependent on a value of the jth attribute, v_{j_k}, of the object τ positions earlier.

Before we describe the method, let us define a *relevant value* for the prediction process to be *an attribute value that is important for determining a certain characteristic of some objects later in a sequence*. As an illustration, let us again consider the sequence of locomotives. By taking a close look at Fig. 1, it is not difficult to see that a relatively large number of locomotives that have four wheels are followed immediately by those that have two windows. For this reason, if a locomotive has four wheels, then the likelihood of its being followed by a locomotive with two windows is greater than that of one or three windows. Hence, the value 'Four' of the attribute NUMBER OF WHEELS can be considered as a relevant value for prediction. It provides helpful information for determining the number of windows of the locomotive in the next position.

By a similar argument, the lack of a medium-length funnel can also be considered as a relevant value for the prediction process. This is because, when compared to those with long and short funnels, relatively few of the locomotives that have medium-length funnels are followed two positions later by

a locomotive that has two windows. Thus, whether a locomotive has four wheels and whether it has a medium funnel provides important information about what characteristics the locomotives one or two positions later in the sequence should possess.

The identification of relevant values for prediction is easy in a deterministic learning environment. If an attribute value, say v_{j_k}, is perfectly correlated with another, v_{i_l}, at a position lag of τ, in the sense that an object characterized by v_{j_k} is always followed at τ positions later by objects characterized by v_{i_l}, then v_{j_k} can be considered a relevant value for prediction. However, as illustrated in our example, the correlation between two attribute values is rarely perfect in the presence of uncertainty. This means that partial, imperfect, yet genuine correlations have to be considered when constructing prediction rules. In other words, as long as the probability for an object that is characterized by the value v_{i_l} given that it is preceded by an object that is characterized by an attribute value, say v_{j_k} (i.e., $Pr(Attr_{i(p+\tau)} = v_{i_l} \mid Attr_{jp} = v_{j_k})$), is it *significantly different* from the probability that an object in the sequence has the characteristic, v_{i_l} (i.e., $Pr(Attr_{i(p+\tau)} = v_{i_l})$), then v_{j_k} can be considered as a relevant value of the prediction process. How great the difference between $Pr(Attr_{i(p+\tau)} = v_{i_l} \mid Attr_{jp} = v_{j_k})$ and $Pr(Attr_{i(p+\tau)} = v_{i_l})$ should be for them to be considered significantly different may be objectively evaluated based on the fact that if they are significantly different, then the difference between the number of objects in the sequence that are characterized by v_{i_l} and are preceded by objects that are characterized by v_{j_k} (i.e., o_{pk}) should deviate significantly from the number of those that are expected (i.e., e_{pk}), to have the characteristic, v_{j_k}, under the assumption that the attribute $Attr_{i(p+\tau)}$ is independent of $Attr_{jp}$.

Since, by simply determining the absolute difference, $\mid o_{pk} - e_{pk} \mid$, we are not provided with any information on the relative degree of the discrepancy, it is necessary to standardize such a difference in some way so as to avoid the influence of the marginal totals. Haberman [16] recommended scaling the difference by computing the *standardized residuals*:

$$z_{lk} = \frac{o_{lk} - e_{lk}}{\sqrt{e_{lk}}}. \tag{2}$$

The standardized residuals have the property that $\sum_{l,k} {z_{lk}}^2$ are distributed asymptotically as chi-square with $(I-1)(J-1)$ degrees of freedom. Also, since z_{lk} is the square root of the X^2 variable, it has an approximate normal distribution with a mean of approximately zero and a variance of approximately one. Therefore, if the absolute value of z_{lk} exceeds 1.96, it would be considered significant at $\alpha = 0.05$ by conventional criteria. We can then conclude that the probability for an object to have the characteristic v_{i_l}, given that it is preceded, at τ positions earlier, by an object with the characteristic, v_{j_k}, is *significantly different* from the probability for an object to have the characteristic v_{i_l}. In other words, v_{j_k} is a relevant value for the prediction process. If an object in the sequence is observed to possess such a characteristic, it is likely that the object at τ positions later in the sequence should be characterized by the value v_{i_l}.

A disadvantage with the use of the standardized residuals is that their approximation to normal distribution is only rough if the asymptotic variance for each z_{lk} is not close enough to 1 [13]. With extra computational cost, one may perform a more precise analysis by using the *adjusted residual* [16] defined as:

$$d_{lk} = \frac{z_{lk}}{\sqrt{\nu_{lk}}}, \tag{3}$$

where ν_{lk} is the maximum likelihood estimate of the variance of z_{lk}, and is given by:

$$\nu_{lk} = (1 - \frac{o_{l+}}{M'})(1 - \frac{o_{+k}}{M'}), \tag{4}$$

where o_{l+} is the total number of objects in the sequence that possess the characteristic v_{i_l} (sum of the lth row in Table I), and o_{+k} is the total number of objects in the sequence that have the characteristic v_{j_k} (sum of the kth column in Table I). An adjusted residual, d_{lk}, provides a better approximation to a standard normal variable than does the standardized residual, z_{lk}, even if the variance of z_{lk} differs substantially from 1.

The analysis of the adjusted residuals can be applied in the same way as with the standardized residuals. Hence, to know which specific values of the ith attribute of an object are genuinely dependent on that of the values of the jth attribute of the object at τ positions earlier, one could first search for unusually large residuals by comparing the absolute value of the adjusted residuals, $d_{lk}, l = 1, 2, \ldots, I,\ k = 1, 2, \ldots, J$ with, 1.96, the 95th percentile of the normal distribution (or 2.35, the 99th percentile, etc. for a greater confidence level). If the absolute value of an adjusted residual, say d_{lk}, is larger than 1.96, we can conclude that the discrepancy between o_{lk} and e_{lk} (i.e., between $Pr(Attr_{i(p+\tau)} = v_{i_l} \mid Attr_{jp} = v_{j_k})$ and $Pr(Attr_{i(p+\tau)} = v_{i_l})$) is significantly different, and therefore v_{j_k} is important for predicting the value of the ith attribute of the object at τ positions later in the sequence.

By noting the sign of the residuals, we can also tell whether it is the presence or the absence of v_{j_k} that is relevant for the prediction process. A d_{lk} that is greater than +1.96 (the 95 percentile of the standard normal distribution) indicates that the presence of v_{j_k} is relevant for predicting the ith attribute of the object at τ positions later. In other words, given that an object is characterized by v_{j_k}, it is more likely for an object to have the value v_{i_l} at τ positions later than for it to have other values. If d_{lk} is smaller than -1.96, it tells us that the absence of v_{j_k} is relevant for prediction in the sense that it is more unlikely for an object characterized by v_{j_k} to be followed at τ positions later in the sequence by an object that has the value v_{i_l}.

Values of $Attr_{jp}$ that show no correlation with any value of $Attr_{i(p+\tau)}$ yield no information about how an object at τ positions later in a sequence should be characterized. Such values are it irrelevant for the prediction process. Their presence introduces noise to the data and could therefore lead to the construction of inaccurate prediction rules. Hence they are disregarded in further analysis.

As an illustration of the above procedure for identifying relevant values for prediction, let us again consider the locomotive problem. As shown above, the result of the chi-square test indicates that the attribute NUMBER OF WINDOWS

TABLE III
THE ADJUSTED RESIDUALS FOR NUMBER OF WINDOWS AND NUMBER OF WHEELS

		NUMBER OF WHEELS		
	z_{lk}			
	ν^2_{lk}			
	d_{lk}	Two	Three	Four
NUMBER OF WINDOWS		-1.00	1.38	-0.15
	Two	0.69	0.71	0.52
		-1.46	1.95	-0.28
		-0.75	-1.67	1.45
	Three	0.67	0.69	0.50
		-1.12	-2.42	2.88
		2.41	0.48	-1.85
	Four	0.78	0.80	0.59
		3.11	0.60	-3.16

of a locomotive is dependent on the attribute NUMBER OF WHEELS of the preceding locomotive. However, based on this test alone, we do not know if a particular locomotive should have a certain number of windows given that the one preceding it has a certain number of wheels. To acquire this information, the adjusted residuals given in Table III can be investigated.

By comparing them with the 5% standard normal deviate, 1.96, we observe that d_{22}, d_{23}, d_{31} and d_{33} are significant. In fact, almost 80% of the total chi-square ($X^2 = 17.82$), is concentrated at these four cells; the other cells contribute very little. Even though one cannot conclusively state that the values 'Three' and 'Four' of the attribute NUMBER OF WHEELS are important for determining if the attribute NUMBER OF WINDOWS in the next locomotive has the value 'Two', and the values 'Two' and 'Four' of NUMBER OF WHEELS are important for determining if the NUMBER OF WINDOWS of the next locomotive has the values 'Three', yet there are obvious reasons for such premises.

Also, from the signs of the deviates, it can be concluded that the presence of the values 'Four' and 'Two' suggests that it is likely for the *NUMBER OF WINDOWS* of the next locomotive to have the values 'Two' and 'Three' (d_{23} and d_{31} are positive) respectively, whereas the presence of the values 'Three' and 'Four' suggests that the next locomotive should not have the values 'Two' and 'Three' respectively (d_{22} and d_{33} are negative).

In other words, a locomotive with four wheels is likely to be followed by one with two windows, while a locomotive with two wheels is likely to be followed by one with three windows. On the other hand, a locomotive with three wheels is unlikely to be followed by one with two windows, and a locomotive with four wheels is unlikely to be followed by one with three windows. Since the absolute values of $d_{1k}, k = 1, 2, 3$ are all less than 1.96, then whether or not a locomotive should have one window cannot be determined based on the number of wheels of the preceding locomotive. That is, the attribute NUMBER OF WHEELS does not provide much information concerning whether a locomotive should have one window.

B. Construction of Prediction Rules Based on Detected Patterns

Since relevant attribute values are important in determining the characteristics of objects later in a sequence, it is necessary to ensure that they are utilized in the construction of prediction rules. A simple way to do this is to represent each detected dependence relation between two attribute values by a rule of the following form:

If $\langle condition \rangle$ then $\langle conclusion \rangle$ with certainty W.

The condition part of the rule specifies the characteristic that an object should possess so that the object at a certain position later in the sequence will take on the attribute value predicted by the conclusion part. In case of the PP problem, since such a prediction cannot usually be made with complete certainty, the amount of certainty involved has to be reflected by the weight W associated with the rule.

As an illustration, suppose that the attribute value v_{i_l} is found to be dependent on v_{j_k} as described in the previous section. The relationship can be represented as:

If $Attr_{jp}$ of an object is v_{j_k} then it is with certainty W that $Attr_{i(p+\tau)}$ of an object located at τ positions later in the sequence has the value v_{i_l},

where $W = W(Attr_{i(p+\tau)} = v_{i_l} / Attr_{i(p+\tau)} \neq v_{i_l} \mid Attr_{j}p = v_{j_k})$ measures the amount of positive or negative evidence provided by v_{j_k} supporting or refuting the object at τ positions later to have the characteristic, v_{i_l}.

Unlike some ad hoc approaches that attempt to mimic some aspects of human reasoning, the derivation of W is based on an information theoretic measure, known as the mutual information, defined between v_{j_k} and v_{i_l} as [25], [34]:

$$I(Attr_{i(p+\tau)} = v_{i_l} : Attr_{jp} = v_{j_k}) = \log \frac{Pr(Attr_{i(p+\tau)} = v_{i_l} \mid Attr_{jp} = v_{j_k})}{Pr(Attr_{i(p+\tau)} = v_{i_l})}. \quad (5)$$

so that $I(Attr_{i(p+\tau)} = v_{i_l} : Attr_{jp} = v_{j_k})$ is positive if and only if $Pr(Attr_{i(p+\tau)} = v_{i_l} \mid Attr_{jp} = v_{j_k}) > Pr(Attr_{i(p+\tau)} = v_{i_l})$ otherwise it is either negative or has

$$
\begin{aligned}
&W(\text{NUMBER OF WINDOWS} = \text{Two}/\text{NUMBER OF WINDOWS} \neq \text{Two} \\
&\qquad \mid \text{NUMBER OF WHEELS} = \text{Four}) \\
&= \log \frac{Pr(\text{NUMBER OF WHEELS} = \text{Four} \mid \text{NUMBER OF WINDOWS} = \text{Two})}{Pr(\text{NUMBER OF WHEELS} = \text{Four} \mid \text{NUMBER OF WINDOWS} \neq \text{Two})} \\
&= 0.94.
\end{aligned}
$$

a value 0. $I(Attr_{i(p+\tau)} = v_{i_l} : Attr_{jp} = v_{j_k})$ measures, intuitively, the increase (if positive) or decrease (if negative) in certainty if the ith attribute of an object is predicted to take on the value v_{i_l} given that the object at τ positions earlier possesses the characteristic, v_{j_k}. Based on the mutual information measure, the weight of evidence for or against a certain prediction of the attribute values of future objects can be assessed quantitatively as follows.

Suppose that v_{i_l} of $Attr_{i(p+\tau)}$ is dependent on v_{j_k} of $Attr_{jp}$. Then the *weight of evidence* provided by v_{j_k} in favor of the ith attribute of the object at τ positions later in the sequence having the value v_{i_l} as opposed to it having some other value can be defined as [25]:

$$
\begin{aligned}
&W(Attr_{i(p+\tau)} = v_{i_l}/Attr_{i(p+\tau)} \neq v_{i_l} \mid Attr_{jp} = v_{j_k}) \\
&= I(Attr_{i(p+\tau)} = v_{i_l} : Attr_{jp} = v_{j_k}) \\
&- I(Attr_{i(p+\tau)} \neq v_{i_l} : Attr_{jp} = v_{j_k}). \qquad (6)
\end{aligned}
$$

In other words, the weight of evidence may be interpreted as a measure of the difference in the gain of information when the ith attribute of an object takes on the value v_{i_l} and when it takes on other values, given that the object that is τ positions in front has the characteristic v_{j_k}. The weight of evidence is positive if v_{j_k} provides positive evidence supporting the ith attribute of the object at τ positions later in the sequence having the value v_{i_l}; otherwise, it is negative. It must be noted that W can also be expressed as:

$$
\begin{aligned}
&W(Attr_{i(p+\tau)} = v_{i_l}/Attr_{i(p+\tau)} \neq v_{i_l} \mid Attr_{jp} = v_{j_k}) \\
&= I(Attr_{i(p+\tau)} = v_{i_l} : Attr_{jp} = v_{j_k}) \\
&- I(Attr_{i(p+\tau)} \neq v_{i_l} : Attr_{jp} = v_{j_k}) \\
&= \log \frac{Pr(Attr_{i(p+\tau)} = v_{i_l} \mid Attr_{jp} = v_{j_k})}{Pr(Attr_{i(p+\tau)} = v_{i_l})} \\
&- \log \frac{Pr(Attr_{i(p+\tau)} \neq v_{i_l} \mid Attr_{jp} = v_{j_k})}{Pr(Attr_{i(p+\tau)} \neq v_{i_l})} \\
&= \log \frac{Pr(Attr_{jp} = v_{j_k} \mid Attr_{i(p+\tau)} = v_{i_l})}{Pr(Attr_{jp} = v_{j_k})} \\
&- \log \frac{Pr(Attr_{jp} = v_{j_k} \mid Attr_{i(p+\tau)} \neq v_{i_l})}{Pr(Attr_{jp} = v_{j_k})} \\
&= \log \frac{Pr(Attr_{jp} = v_{j_k} \mid Attr_{i(p+\tau)} = v_{i_l})}{Pr(Attr_{jp} = v_{j_k} \mid Attr_{i(p+\tau)} \neq v_{i_l})} \qquad (7)
\end{aligned}
$$

The prediction rules that are constructed by the above procedure describe the object-generating process probabilistically. As an illustration of the rule-construction procedure, let us return to the problem of predicting locomotives. Since the value 'Four' of the attribute NUMBER OF WHEELS is correlated with the value 'Two' of the attribute NUMBER OF WINDOWS for the next locomotive, the weight of evidence provided by the former in favour of the latter as opposed to other values is (see above):

In other words, the following rule can be constructed:

- If a locomotive has four wheels then it is with certainty 0.94 that the locomotive located at one position later in the sequence has two windows.

Based on the other relevant values, all the rules can be constructed. They are:

1) If a locomotive has four wheels then it is with certainty 0.94 that the locomotive located at one position later in the sequence has two windows.
2) If a locomotive has three wheels then it is with certainty -2.82 that the locomotive located at one position later in the sequence has two windows.
3) If a locomotive has two wheels then it is with certainty 1.94 that the locomotive located at one position later in the sequence has three windows.
4) If a locomotive has four wheels then it is with certainty -2.02 that the locomotive located at one position later in the sequence has three windows.
5) If a locomotive has four wheels then it is with certainty 0.91 that the locomotive located at two position later in the sequence has three windows.
6) If a locomotive has a medium-length funnel then it is with certainty 1.25 that the locomotive located at two position later in the sequence has one window.
7) If a locomotive has a medium-length funnel then it is with certainty -1.44 that the locomotive located at two position later in the sequence has two windows.
8) If a locomotive has two stripes then it is with certainty 1.03 that the locomotive located at one position later in the sequence has one window.
9) If a locomotive has two stripes then it is with certainty -1.01 that the locomotive located at one position later in the sequence has two windows.
10) If a locomotive has one window then it is with certainty -1.38 that the locomotive located at one position later in the sequence has one window.
11) If a locomotive has one window then it is with certainty 1.20 that the locomotive located at one position later in the sequence has two windows.
12) If a locomotive has two windows then it is with certainty -1.42 that the locomotive located at one position later in the sequence has two windows.

$$
\begin{aligned}
&W(Attr_{i(M+h)} = val_{i(M+h)} / Attr_{i(M+h)} \neq val_{i(M+h)} \\
&\quad \mid Attr_{jp} = val_{jp}; j = 1, \ldots, n; p = M, M-1, \ldots, (M-L)+1) \\
&= W(val_{i(M+h)} / \overline{val}_{i(M+h)} \mid \overline{val}_{jp}; j = 1, \ldots, n; p = M, M-1, ..., (M-L)+1) \\
&= \log \frac{Pr(val_{i(M+h)} \mid val_{jp}; j = 1, \ldots, n; p = M, M-1, \ldots, (M-L)+1)}{Pr(val_{i(M+h)})} \\
&\quad - \log \frac{Pr(\overline{val}_{i(M+h)} \mid \overline{val}_{jp}, j = 1, \ldots, n; p = M, M-1, \ldots, (M-L)+1)}{Pr(\overline{val}_{i(M+h)})} \\
&= \log \frac{Pr(val_{jp}; j = 1, \ldots, n; p = M, M-1, \ldots, (M-L)+1 \mid val_{i(M+h)})}{Pr(\overline{val}_{jp}; j = 1, \ldots, n; p = M, M-1, \ldots, (M-L)+1 \mid val_{i(M+h)})}
\end{aligned} \tag{8}
$$

It must be noted that a negative weight of evidence, for example, in Rule 2 above implies that if a locomotive has three wheels, then there is negative evidence against the next locomotive having two windows. In other words, it is more likely for such a locomotive to have one or three windows than two.

C. Prediction of Future Objects

Given a set of prediction rules which were constructed based on the detected patterns inherent in a sequence of objects, the characteristics of a future object may be predicted based on them. To illustrate how such predictions can be made, let us suppose, once again, that we are given a sequence S of M objects, $obj_1, \ldots, obj_M$, and suppose that we are to predict the value of the ith attribute $Attr_{i(M+h)}$ of the object obj_{M+h} which is h positions behind the most recently observed one in S, obj_M. To determine if the value of $Attr_{i(M+h)}$ is dependent on the objects in S, there is a need to know which attributes of which objects may affect its value. Assuming that the objects are generated probabilistically in such a way that the characteristics of an object at a certain position depend solely on that of a maximum of L objects before it, the prediction process begins by searching through the space of prediction rules to determine how the characteristics of $obj_M, obj_{M-1}, \ldots, obj_{(M-L)+1}$ may affect the value of $Attr_{i(p+\tau)}$ of obj_{M+h}.

This search process proceeds by matching the attribute values val_{jp} (where $j = 1, 2, \ldots, n$ and $p = M, M-1, \ldots, (M-L)+1$), of the objects $obj_M, obj_M - 1, \ldots, obj_{(M-L)+1}$, against the subset of prediction rules whose conclusion parts predict what values the ith attribute of an object at $h, h+1, \ldots, (h+L)-1$ positions later will take on. An attribute value that satisfies the condition part of a rule in such a subset, therefore, affects the value of the ith attribute of the object at $M+h$. Hence, this value provides a certain amount of evidence, reflected by the weight of the rule, supporting or refuting the ith attribute to take on the value predicted by the conclusion part.

In the presence of uncertainty, it is possible for the ith attribute to be predicted to take on several different values based on different attributes of different objects in S. Suppose in our problem that we are to predict the number of windows that the locomotive at position 60 will take on based on the last two at positions 58 and 59. These two locomotives have characteristics of a medium-length funnel, one stripe, three wheels, and one window and a long funnel, two stripes, four wheels, and one window, respectively.

- Since the locomotive at position 58 has a medium-length funnel, the locomotive at position 60 should have one window according to Rule 6.
- Since the locomotive at position 58 has a medium-length funnel, the locomotive at position 60 should not have two windows according to Rule 7.
- Since the locomotive at position 59 has two stripes, the locomotive at position 60 should have one window according to Rule 8.
- Since the locomotive at position 59 has two stripes, the locomotive at position 60 should not have two windows according to Rule 9.
- Since the locomotive at position 59 has four wheels, the locomotive at position 60 should have two windows according to Rule 1.
- Since the locomotive at position 59 has four wheels, the locomotive at position 60 should not have three windows according to Rule 4.
- Since the locomotive at position 59 has one window, the locomotive at position 60 should not have one window according to Rule 10.
- Since the locomotive at position 59 has one window, the locomotive at position 60 should have two windows according to Rule 11.

In summary, there is, therefore, both positive and negative evidence for the locomotive at position 60 to have one or two windows.

In order to decide what specific value $Attr_{i(M+h)}$ is most likely to take on, it is noted that the attribute values of the objects $obj_M, obj_{M-1}, \ldots, obj_{(M-L)+1}$ that match the prediction rules can be considered as providing some evidence for or against $Attr_{i(M+h)}$ to have a certain value. For this reason, such a decision can be made by combining the various items of evidence.

To quantitatively estimate and combine the evidence so that they can be compared, a measure of evidence is proposed here. It has the property that its value increases with the number and

$$W(Attr_{i(M+h)} = v_{i_l} / Attr_{i(M+h)} \neq v_{i_l} \mid val_{[1]}, \ldots, val_{[m]}) > W(Attr_{i(M+h)} = v_{i_q} / Attr_{i(M+h)} = v_{i_q} \mid val_{[1]}, \ldots, val_{[m]}), \quad q = 1, 2, \ldots, I' \text{and} q \neq l \tag{11}$$

strength of the positive items of evidence supporting a specific value of $Attr_{i(M+h)}$, and decreases with the number and the strength of the negative items of evidence refuting such a value. This measure, known also as the *weight of evidence* measure measures the evidence, provided by the attribute values val_{jp}, $j = 1, 2, \ldots, n$ and $p = M, M-1, \ldots, (M-L)+1$ of the objects $obj_M, obj_{M-1}, \ldots, obj_{(M-L)+1}$, in favor of $Attr_{i(M+h)}$ taking on a certain value. It is defined as (see (8) on the previous page):

Suppose that, of all the characteristics of the objects, $obj_M, obj_{M-1}, \ldots, obj_{(M-L)+1}$, only m of them, $val_{[1]}, \ldots, val_{[j]}, \ldots, val_{[m]}, val_{[j]} \in \{val_{jp} \mid j = 1, \ldots, n; p = M, M-1, \ldots, (M-L)+1\}$, are found to match one or more prediction rules. Then the weight of evidence can be simplified into [5]:

$$W(val_{i(M+h)} / \overline{val}_{i(M+h)} \mid val_{jp}; j = 1, 2, \ldots, n; p = M, M-1, \ldots, (M-L)+1) = W(val_{i(M+h)} / \overline{val}_{i(M+h)} \mid val_{[1]}, \ldots, val_{[m]}) \tag{9}$$

If there is no a priori knowledge concerning the interrelation of the attribute values in the problem domain, then the weight of evidence provided by the attribute values of $obj_M, obj_{M-1}, \ldots, obj_{(M-L)+1}$ in favor of $Attr_{i(M+h)}$ taking the value $val_{i(M+h)}$ as opposed to its taking other values is equal to the sum of the weights of evidence provided by each individual attribute value of the objects $obj_M, obj_{M-1}, \ldots, obj_{(M-L)+1}$ that is relevant for the prediction task. For this reason, we can write [5]:

$$W(val_{i(M+h)} / \overline{val}_{i(M+h)} \mid val_{[1]}, \ldots, val_{[m]}) = \sum_{j=1}^{m} W(val_{i(M+h)} / \overline{val}_{i(M+h)} \mid val_{[j]}). \tag{10}$$

Hence, intuitively, the total amount of evidence supporting or refuting $Attr_{i(M+h)}$ to take on the value $val_{i(M+h)}$ is equal to the sum of the individual pieces of evidence provided by each relevant attribute values of the objects $obj_M, obj_{M-1}, \ldots, obj_{(M-L)+1}$ for or against such a value.

In brief, the strategy for predicting a certain characteristic of a future object based on a sequence of M objects can be summarized as follows. To predict the ith attribute of an object $obj_{(M+h)}$ based on the attributes of the L most recently observed ones, obj_M, obj_{M-1}, $\ldots$, $obj_{(M-L)+1}$, the set of prediction rules is searched to determine which characteristics of these L objects affect the value of the ith attribute, $Attr_{i(M+h)}$ of $obj_{(M+h)}$. This search process proceeds by matching each attribute value of these objects against the rules whose condition parts predict an object to take on a certain value of the ith attribute at $h, h+1, \ldots, (h+L)-1$ positions later. Therefore, if an attribute value of one of the L objects has satisfied the condition part of a rule, then it affects the value of the ith attribute of the object $obj_{(M+h)}$ at a later position specified by the rule. The attribute value can be considered to provide some evidence for or against $Attr_{i(M+h)}$ taking on the value predicted by the conclusion part of it. The strength of this evidence is given by the weight of the rules. Since the attributes of the objects $obj_M, obj_{M-1}, \ldots, obj_{(M-L)+1}$ may or may not provide evidence, and since even those that do may support different values, these evidences are quantitatively measured and combined for comparison in order to find the most probable value that $Attr_{i(M+h)}$ should take on. Based on the weight of evidence measure, $Attr_{i(M+h)}$ is predicted to take on the value v_{i_l}, if (see (11) above)
where $I'(<= I)$ denotes the number of values of the ith attribute that are dependent on the attributes of the L most recent objects in the sequence S. It should be noted that it is possible for two different plausible values to have the same greatest weight of evidence. In this case, there may be more than one plausible value that $Attr_{i(M+h)}$ may take on. Furthermore, if there is no evidence for or against any specific value of the ith attribute, prediction will be refrained. Instead making a random guess, $Attr_{i(M+h)}$ can be given the value which the majority of the objects in the sequence have. If it happens that there is no relevant value for determining the future value of the ith attribute of $obj_{(M+h)}$, then either the generation of the sequence is completely nondeterministic, or there are insufficient generated objects for any prediction to be made.

As an illustration of the above procedure for determining the attribute of a future object, suppose we are interested in predicting the number of windows that the locomotive at position 60 may have.

- Since the locomotive at position 58 have a medium-length funnel, the positive evidence for the locomotive at position 60 to have one window is 1.25 according to Rule 6;
- Since the locomotive at position 58 has a medium-length funnel, the negative evidence the locomotive at position 60 to have two windows is -1.44 according to Rule 7;
- Since the locomotive at position 59 has two stripes, the positive evidence for the locomotive at position 60 to have one windows is 1.03 according to Rule 8;
- Since the locomotive at position 59 has two stripes, the negative evidence against the locomotive at position 60 to have two windows is -1.01 according to Rule 9;

$$W(\text{NUMBER OF WINDOWS} = \text{One}/\text{NUMBER OF WINDOWS} \neq \text{One}$$
$$\mid \text{medium-length funnel, one stripes, three wheels, one window,}$$
$$\text{long funnel, two stripes, four wheels , one window})$$
$$= W(\text{NUMBER OF WINDOWS} = \text{One}/\text{NUMBER OF WINDOWS} \neq \text{One}$$
$$\mid \text{medium-length funnel , two stripes , one window})$$
$$= 0.90.$$

Similarly:

$$W(\text{NUMBER OF WINDOWS} = \text{Two}/\text{NUMBER OF WINDOWS} \neq \text{Two}$$
$$\mid \text{medium-length funnel , one stripes , three wheels , one window , ,}$$
$$\text{long funnel , two stripes , four wheels, one window})$$
$$= W(\text{NUMEBR OF WINDOWS} = \text{Two}/\text{NUMBER OF WINDOWS}$$
$$\neq \text{Two} \mid \text{medium-length funnel , two stripes , four wheels , one window})$$
$$= -0.31.$$

$$W(\text{NUMBER OF WINDOWS} = \text{Three}/\text{NUMBER OF WINDOWS} \neq \text{Three}$$
$$\mid \text{medium-length funnel , one stripes , three wheels , one window ,}$$
$$\text{long funnel , two stripes, four wheels , one window})$$
$$= W(\text{NUMEBR OF WINDOWS} = \text{Three}/\text{NUMBER OF WINDOWS} \neq \text{Three}$$
$$\mid \text{four wheels})$$
$$= -2.02.$$

- Since the locomotive at position 59 has four wheels, the positive evidence for the locomotive at position 60 to have two windows is 0.94 according to Rule 1;
- Since the locomotive at position 59 has four wheels, the negative evidence against the locomotive at position 60 to have three windows is -2.02 according to Rule 4;
- Since the locomotive at position 59 has one window, the negative evidence against the locomotive at position 60 to have one windows is -1.38 according to Rule 10;
- Since the locomotive at position 59 has one window, the positive evidence for the locomotive at position 60 to have two windows is 1.21 according to Rule 12.

Based on these rules, it seems that there is both positive and negative evidence supporting or refuting locomotive number 60 to have one or two windows. The weight of evidence for or against the attribute NUMBER OF WINDOWS taking on different values can be computed as follows (see above):

Since the weight of evidence for the value 'One' as opposed to the values 'Two' and 'Three' is the greatest, we predict, therefore, locomotive number 60 to have one window. There is negative evidence against its having two or three windows.

V. Performance Evaluation

The proposed inductive learning method for solving the PP problem has been implemented in a system called OBSERVER-II. This system has been tested with both simulated and real-world prediction problems. In this section, the results of these tests and how OBSERVER-II is able to successfully solve them while other systems fail to do so are discussed.

A. An Experiment Involving a Sequence of Movements

A hypothetical scenario is presented here, which demonstrates the capability of the proposed method to discover the rules governing the interrelations of the movements of a set of objects over a period of time. Suppose four objects 1, 2, 3, and 4 are moving around in a bounded area and OBSERVER-II is taking a snapshot of them at fixed time intervals (Fig. 2). Suppose also that each of these objects makes one move during every time interval.

In order to simulate the PP problem, the data for this experiment is generated stochastically in such a way that: 1 moves randomly; 2 always follows 1 (at a time lag of 1) by moving in the same direction as 1; 3 moves in the same direction as 2 about 55% of the time whenever the previous move of 2 was to E, S, or W, but if 2 moves to the N, then 3 may move in any direction except N. 4 moves in a direction that is opposite to that of 3 whenever 3 moved in the same direction as 2; otherwise, it moves either to E or W.

To detect these probabilistic patterns, it must be noted that an observation made at time t can be described by four attributes, $Attr_{jt}, j = 1, \ldots, 4$ corresponding to the movements of the four objects, 1, 2, 3 and 4, respectively, so that $domain(Attr_{jt}) = \{E, S, W, N\}$. To determine the future position of an object, say 3, we have to know how it

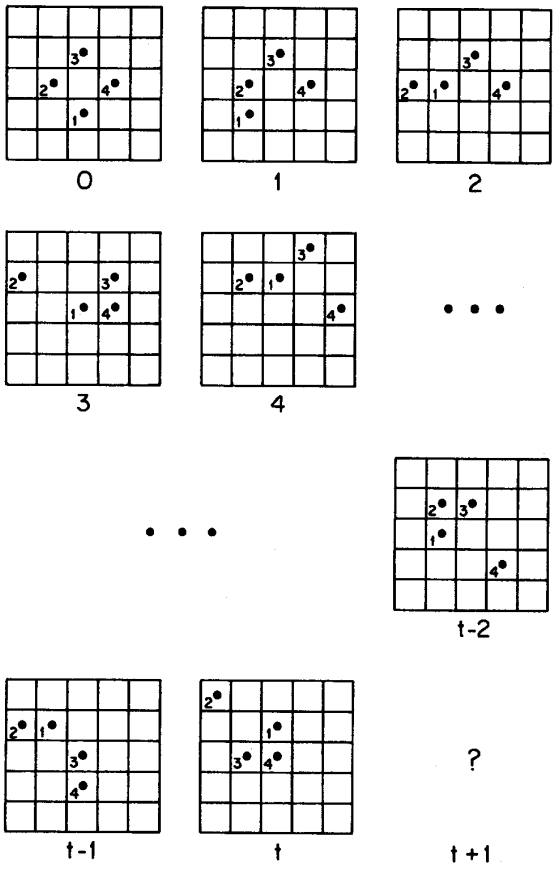

Fig. 2. A sequence of snapshots showing the positions of four objects.

moves around in the enclosed area, and whether the movement of any other object, say 2, affects its movement or not. In other words, we have to determine if the movement of 3 is genuinely dependent on the movement of 2 at a certain time unit earlier.

By applying the chi-square test, we discover that there is a significant dependence relationship between the movements of 3 and that of 2 at a time difference of one unit ($X^2_{9,0.95}$) (Table IV).

However, based on this test alone, how the movement of 3 is dependent on the last movement of 2 is not completely clear. A more effective investigation, as described in the last section, has to be conducted. We thus analyze the adjusted residuals given in Table V, to determine which specific movement of 3 is genuinely dependent on that of 2 at one time unit earlier.

By comparing them with the 5% standard normal deviate, 1.96, we observe that d_{EE}, d_{SS}, d_{WW}, and d_{NN} are significant. In fact, about 60% of the total X^2 is concentrated at these five cells; the other cells contribute very little. Even though one cannot, based on, say d_{EE}, conclusively state that 3 moves to E whenever 2 has made its last move towards E, there are obvious reasons for such premises.

Also, from the signs of the deviates, we can conclude that the dependences of E, S, W of 3 on E, S, W of

TABLE IV
A 4 × 4 CONTINGENCY TABLE FOR $Attr_{2t}$ AND $Attr_{3(t+1)}$

		$Attr_{2t}$				
	o_{pk} (e_{pk})	E	S	W	N	TOTAL
$Attr_{3(t+1)}$	E	11 (6.49)	5 (7.18)	3 (5.47)	8 (7.86)	27
	S	3 (5.77)	11 (6.38)	2 (4.86)	8 (6.98)	24
	W	2 (4.57)	2 (5.05)	8 (3.85)	7 (5.53)	19
	N	3 (2.17)	3 (2.39)	3 (1.82)	0 (2.62)	9
	TOTAL	19	21	16	23	79

2, respectively, are positive, whereas N of 3 is negatively dependent on N of 2. This suggests that the movement of 2 to E, S, or W provides evidence supporting that the next movement of 3 will be towards E, S or W respectively; whereas the movement of 2 to N implies that there is negative evidence against the next movement of 3 to be to N. In a similar way, other dependence relationships between the movements of different objects at different times are detected, and the probabilistic patterns by which the data are generated are successfully discovered.

Based on the detected relationships in the observed movements of the four objects, a set of prediction rules can be constructed. For example:

- If the movement of 2 is towards W then it is with certainty 1.66 that the next movement of 3 is to W

where 1.66 is the weight of evidence for 3 to move to W as opposed to its moving to E, S, or N, provided that the last movement of 2 is also towards W.

To illustrate how these rules can be employed for prediction of the future movements of the four objects, let us suppose that we are to predict the next movement of object 3, given the sequence in Fig. 2.

To do this, we have to know which movements of 3 are dependent on the previous movements of other objects. (In this experiment, the maximum lookback L is taken to be 1.) From a search through the set of prediction rules, we find that the movement of 3 is independent of the last movement of all others with the exception of 2 in the way described above. Therefore, to compute the weight of evidence in favour of 3

TABLE V
THE ADJUSTED RESIDUALS FOR $Attr_{2t}$ AND $Attr_{3(t+1)}$

		$Attr_{2t}$			
	z_{lk}				
	ν_{lk}^2				
	d_{lk}	E	S	W	N
CLASS		1.77	-0.81	-1.06	0.05
	E	0.71	0.70	0.73	0.68
		2.50	-1.17	-1.46	0.07
		-1.15	1.83	-1.30	0.38
	S	0.73	0.72	0.75	0.70
		-1.59	2.56	-1.74	0.54
		-1.20	-1.36	2.12	0.62
	W	0.76	0.75	0.78	0.73
		-1.58	-1.82	2.72	0.85
		0.57	0.39	0.87	-1.62
	N	0.82	0.81	0.84	0.79
		0.69	0.49	1.04	-2.04

moving in a particular direction, say W, we compute:

$$W(Attr_{3(t+1)} = W/Attr_{3(t+1)} \neq W \mid Attr_{1t} = E, \; Attr_{2t} = W, Attr_{3t} = W, Attr_{4t} = W) = W(Attr_{3(t+1)} = W/Attr_{3(t+1)} \neq W \mid Attr_{2t} = W) = 1.66.$$

By a similar procedure, the total weight of evidence in favour of 3 moving E, S, and N as opposed to its moving in other directions can be determined, and 3 is predicted to move to W, which is the direction supported by the strongest evidence. The next movements of all the other objects are also correctly predicted with the exception of 1: no pattern can be detected concerning its movement as it moves independently of the others in a completely random manner. It should be noted that the movements of objects 2, 3, and 4 are correctly predicted even though the relative positions they occupied in the 5×5 matrix have, in some instances, never been observed before. Since this PP problem involves data that are generated probabilistically, SPARC and the other programs described in Section 2 are not able to solve it.

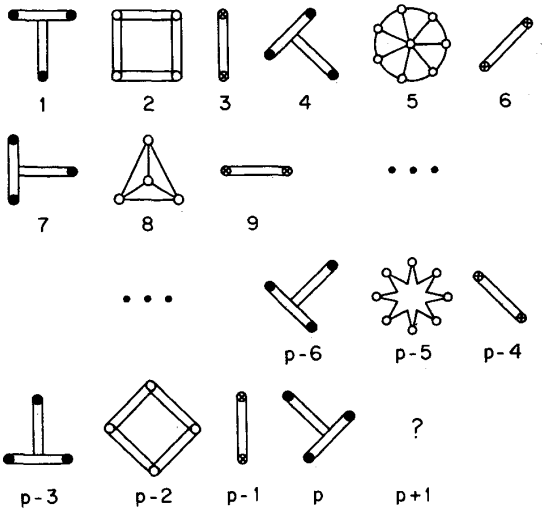

Fig. 3. A sequence of geometric figures.

B. An Experiment Involving a Sequence of Geometric Figures

Suppose that the OBSERVER-II is given snapshots of an ongoing process that generates a sequence of geometric figures (Fig. 3) [10]. Suppose also that each of the figures is described by four attributes: SHAPE, NUMBER OF NODES, NODE TEXTURE, and ORIENTATION. The problem is to predict the characteristics of the figure at $p + 1$.

Based on the learning algorithm described in the last section, the OBSERVER-II is able to discover the patterns underlying the sequence of geometric figures. The figures in the sequence can be grouped together into subsequences of triplets so that the nodes of the figures in each subsequence have textures in the order 'solid black', 'blank', 'cross'; and the corresponding shapes are always 'T-junction', 'an object with 4 or 8 nodes', 'bar'. The orientation of the T-junction changes by -45 degrees each time with respect to its last appearance, whereas that of the bar changes by +45 degrees in a similar manner. The number of nodes of the middle figure in each triplet alternates between 4 and 8.

By detecting these patterns, it is predicted that the figure at position $p + 1$ has 8 blank nodes and that its shape is neither a T-junction nor a bar (i.e. there is negative evidence against its taking either of these two shapes). Since the shape of the figure cannot be determined, its orientation with respect to its last appearance is not predicted also.

It must be noted that the maximum lookback is taken to be 1 in this experiment. However, all the predictions can still be

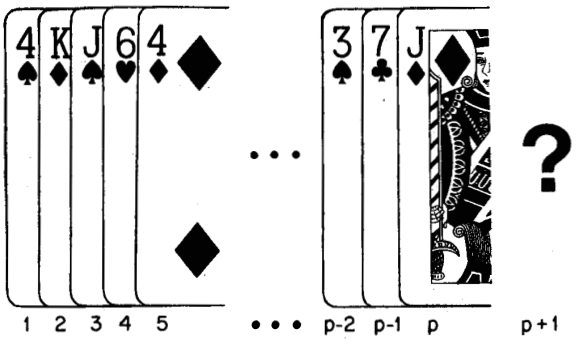

Fig. 4. A sequence of cards (the mainline) in a game of Eleusis.

made accurately. The reason why this is possible may not be immediately obvious to human observers, who would normally predict the number of nodes of the figure at $p+1$ by looking at those of the figure at position $p-2$. OBSERVER-II, however, is capable of discovering that the number of nodes of a figure can be predicted, with complete certainty, based on the shape and orientation of the last figure alone. It discovers that if the last figure is a T-junction at +135 degrees, the next figure will definitely have 8 nodes. Such a discovery is totally unexpected before the experiment.

The SPARC program has also been tested with this experiment. However, it is unable to correctly predict the characteristics of the next figure, due to its inability to form composite models. And as a result, it is not able to find a sequence-generating rule in a sequence of objects that involves nested periodic structures [10].

C. An Experiment Involving a Sequence of Playing Cards

Suppose that OBSERVER-II is given a layout (the mainline) of a game of Eleusis (Fig. 4) [10]. Suppose also that each card of the sequence is described by the attributes: SUIT, RANK, COLOR, FACEDNESS, PARITY, PRIMENESS and the attributes REL_SUIT, DIFF_RANK where REL_SUIT takes on the value of the relation $(>, <, =)$ between the suits of two adjacent cards and DIFF_RANK takes on the value of the difference in suit modulo 4 between their ranks. (These attributes can either be explicitly represented or they can be derived by the user providing the system with a definition for each of them.) The problem is to find the secret rule that governs the layout of the cards, and to predict the next card in the sequence.

Based on the learning algorithm described in the last section, OBSERVER-II is able to discover the secret rule that states that: if the rank of the card is higher than or equal to that of the previous one, its suit will also be one suit higher (modulo 4); otherwise, if the rank of the card is lower than or equal to that of the previous one, then its suit will be three suits higher (modulo 4). Based on these rules, the next card in the sequence is predicted to have a rank lower than that of J, and the suit of the card will be 'club' since the last suit of the last card is a 'diamond'. As expected, the exact rank of the card is not predicted, due to a lack of evidence to support any specific value.

The results of this experiment show that OBSERVER-II is able to discover rules that are not immediately obvious to a human. It performs better than the SPARC program, which is unable to discover the secret rule of this game due to the fact that none of the models it considers fit the data [10].

D. An Experiment Involving a Sequence of Weather Records

This experiment is performed on a set of real-life data involving the mean temperature of twelve European cities: Copenhagen, Edinburgh, Geneva, Stockholm, London, Rome, Marseilles, Milan, Paris, Berlin, Vienna, and Oslo. The data are taken from [23], and they represent the mean temperature for the month of July collected over a period of 175 years (1751-1975). In other words, a sequence of 175 weather records, each of which is described by a total of 12 attributes, is available. OBSERVER-II's task is to discover if there is any underlying pattern in the data, and to forecast future July temperature based on the detected patterns.

Since it is more meaningful, especially in long-term forecasting, for a mean temperature to be predicted to be within a certain interval rather than at a certain point, the set of possible values that an attribute may take on is divided into a few intervals. Instead of finding these intervals arbitrarily, and thereby losing information, a procedure based on the maximum entropy formalism is employed so that the probability for an observed temperature to fall into any interval is approximately the same [37]. This procedure allows the temperature scale to be divided into different intervals, while at the same time it reduces the loss of information to a minimum. By such a procedure, the recorded mean temperatures for each city were grouped into four intervals, as shown in Table VI.

By applying the method proposed in this paper, OBSERVER-II is able to discover how the temperatures of the different cities affect each other at different periods of time. The discovered hidden regularities in the data are then represented by a set of prediction rules, whose performance is evaluated by employing some of the available weather records for testing. The mean temperature of each of the 12 cities for 27 years (from 1949 to 1975) is predicted based on the weather records of the previous years (the maximum lookback is taken to be 6). The performance of the prediction rules is then compared to that of random guessing and to that of guessing based on the most numerous value in the data (i.e., the temperature is predicted to be in the interval into which most of the other observations fall). The results for the test and how different methods compare with each other are shown in Table VII.

It should be noted that the predictions made by OBSERVER-II are better than those made by the other methods. For the cities of Stockholm, Rome, Milan, Berlin, and Oslo, the percentages of accurate predictions are all higher than 74%. It must also be noted that the mean temperatures for some cities – Geneva, London, and Marseilles – are difficult to predict accurately. Their temperatures seem to be independent of that of other cities considered here. In

TABLE VI
RANGES OF TEMPERATURE (IN CELCIUS) REPRESENTED BY EACH INTERVAL

		Interval Ranges			
		1	2	3	4
	Copenhagen	12.6-16.0	16.1-17.2	17.3-17.9	18.0-21.4
	Edinburgh	12.1-14.0	14.1-14.6	14.7-15.3	15.4-18.4
	Geneva	15.6-18.3	18.4-19.2	19.3-20.5	20.6-23.4
	Stockholm	12.9-16.0	16.1-17.0	17.1-18.4	18.5-21.4
	London	14.2-16.0	16.1-16.7	16.8-18.1	18.2-20.5
city	Rome	20.5-23.6	23.7-24.4	24.5-25.2	25.3-27.6
	Marseilles	19.8-21.8	21.9-22.6	22.7-23.3	23.4-25.6
	Milan	18.9-22.9	23.0-23.7	23.8-25.0	25.1-28.3
	Paris	15.2-17.4	17.5-18.3	18.4-19.7	19.8-22.0
	Berlin	5.4- 7.6	7.7- 8.9	9.0- 9.9	10.0-13.6
	Vienna	4.3- 8.9	9.0- 9.8	9.9-10.9	11.0-14.6
	Oslo	3.1- 6.0	6.1- 6.9	7.0- 7.9	8.0-12.7

TABLE VII
COMPARISON OF THREE FORECASTING METHODS

	Random Guess		Simple Majority		OBSERVER-II		
Rates (%)	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	No Prediction
Copenhagen	25.0	75.0	25.9	74.1	29.6	66.7	3.7
Edinburgh	25.0	75.0	22.2	77.8	59.3	29.6	11.1
Geneva	25.0	75.0	18.5	81.5	18.5	77.8	3.7
Stockholm	25.0	75.0	29.6	70.4	81.5	14.8	3.7
London	25.0	75.0	29.6	70.4	22.2	70.4	7.4
Rome	25.0	75.0	37.0	63.0	81.5	14.8	3.7
Marseilles	25.0	75.0	37.0	63.0	18.5	63.0	18.5
Milan	25.0	75.0	7.4	92.6	88.9	11.1	0.0
Paris	25.0	75.0	18.5	81.5	29.6	59.3	11.1
Berlin	25.0	75.0	25.9	74.1	74.1	22.2	3.7
Vienna	25.0	75.0	40.7	59.3	25.9	51.9	22.2
Oslo	25.0	75.0	37.0	63.0	74.1	14.8	11.1

view of the difficulty in long-term weather forecasting, the experimental results are very encouraging.

VI. CONCLUSION

Considering that decision-making about the likelihood of some future events based on past information plays a crucial role in scientific progress as well as in everyday life and that there is a growing interest in the development of expert systems for prediction tasks in many fields, it is important that an efficient strategy be developed to assist with predictions. Existing methods are not suitable for dealing with prediction problems in the presence of uncertainty. For example, traditional cognitive model-based approaches can only deal with letter- or number-series extrapolation problems and they cannot handle noisy data. More recent AI method, such as the SPARC program, handles the prediction problem by adopting a model-driven learning strategy. The sequence generating process, in such case, has to be assumed to fit certain types of rule models. Furthermore, the sequence also has to be assumed to be transformable, by at least one of several different functions, into another. The need for such assumptions makes it difficult for SPARC to tolerate noisy data.

To deal with uncertainty in prediction tasks, we propose an inductive learning method in this paper. The proposed method is based on a new probabilistic inference technique and is thus able to discover patterns in sequences in which the characteristics of an object are not completely dependent on those preceding it. This method has been implemented in the OBSERVER-II system and was tested with different sets of real and simulated data. The experimental results have demonstrated the capability of the system in solving PP problems in which other existing prediction systems are not designed to.

In conclusion, OBSERVER-II is able to discover the probabilistic patterns inherent in a sequence of objects and to construct, with or without supervision, prediction rules based on these patterns. It can be used to solve complex real-world problems where predictions have to be made in the presence of uncertainty and a probabilistic answer based on the previous observations is more appropriate than an exact one. OBSERVER-II has the capability to discover hidden patterns and to explain the behaviour of certain sequence-generating processes and causal relationships that a user is not likely to be immediately aware of or to fully understand. It represents an important step towards the goal of automating the knowledge acquisition process in the construction of knowledge-based systems for applications involving forecasting of future events.

REFERENCES

[1] G. Allen, W. Bolam and V. Ciesielski, "Evaluation of an expert system to forecast rain in Melbourne," in *Proc. of the First Australian Artificial Intelligence Conference* (Melbourne, Australia), 1986.
[2] K. C. C. Chan and A. K. C. Wong, "PIS: A Probabilistic inference system," in *Proc. of the 9th International Conference on Pattern Recognition* (Rome, Italy), pp. 360-364, 1988.
[3] K. C. C. Chan, A. K. C. Wong and D.K.Y. Chiu, "Discovery of probabilistic rules for prediction," in *Proc. of the Fifth IEEE Conference on Artificial Intelligence Applications* (Miami, Florida), pp. 223-229, 1989.
[4] K. C. C. Chan, and A. K. C. Wong, "A Statistical technique for extracting classificatory knowledge from databases," in *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro and W.J. Frawley, Eds., Cambridge, MA:MIT Press, pp.108-123, 1991.
[5] K. C. C. Chan, "Inductive learning in the presence of uncertainty," Ph.D. Dissertation, Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, 1989.
[6] K. C. C. Chan, and A. K. C. Wong, "APACS: A System for automated pattern analysis and classification," *Computational Intelligence*, vol. 6, pp. 119-131, 1990.

[7] B. M. Charpin, "PANISSE: A Prototype expert system to forecast French France/U.S. dollar exchange rate," in *Expert Systems and Knowledge Engineering*, T. Bernold, Ed., Amsterdam, Holland:Elsvier Science, 1986.
[8] J. De Kleer, and J.S. Brown, "A Qualitative physics based on confluences," *Artificial Intelligence*, vol. 24, pp. 7-83, 1984.
[9] T. G. Dietterich and R.S. Michalski, "Learning and generalization of characteristic descriptions: evaluation criteria and comparative review of selected methods," in *Proc. of the International conference on Artificial Intelligence* (Vancouver, B.C.), pp. 223-231, 1979.
[10] T. G. Dietterich and R.S. Michalski, "Discovering patterns in sequences of events", *Artificial Intelligence*, vol. 25, pp. 187-232, 1985.
[11] R. Elio and J. De Haan, "Representing quantitative and qualitative knowledge in a knowledge-based storm-forecasting system," *International Journal of Man-Machine Studies*, vol. 25, pp. 523-547, 1986.
[12] G. W. Ernst and A. Newell, *GPS: A Case Study in Generality and Problem Solving*. New York, NY:Academic Press, 1969.
[13] B. Fingleton, *Models of Category Counts*, Cambridge, England: Cambridge University Press, 1984.
[14] K. Forbus, "Qualitative Process Theory", *Artificial Intelligence*, vol. 24, pp. 85-168, 1984.
[15] E. Fredkin, "Techniques using LISP for automatically discovering interesting relations in data," in *The programming language LISP*, E. C. Berkeley and D. Bobrow, Eds., Cambridge, MA: Information International, 1964.
[16] Haberman, S. J. "The analysis of residuals in cross-classified tables," *Biometrics*, vol. 29, pp. 205-220, 1973.
[17] T. G. Holzman, J. W. Pellegrino and R. Zlaser, "Cognitive variables in series completion," *Journal of Educational Psychology*, vol. 75, no. 4, pp. 603-618, 1983.
[18] H. H. Hsu, "Computerized technology, information management, and commodity forecasting system for agriculture," in *Computer Vision, Image Processing and Communications Systems and Applications*, P.S.P. Wang, Ed., Philadelphia, PA: World Scientific, pp. 72-78, 1986.
[19] K. Kotovsky and H. A. Simon, "Empirical tests of a theory of human acquisition of concepts for sequential events," *Cognitive Psychology*, vol. 4, pp. 399-424, 1973.
[20] J. LeFevre and J. Bisanz, "A Cognitive analysis of number-series problems: sources of individual differences in performance," *Memory and Cognition*, vol. 14, no. 4, pp. 287-298, 1986.
[21] F. H. Merrem, "Two expert systems used in weather forecasting," in *Proceedings IEEE Expert Systems in Government Symposium* (McLean, VA), pp. 342-343, 1986.
[22] R. S. Michalski, H. Ko, and K. Chen, "Qualitative prediction: The SPARC/G methodology for inductively describing and predicting discrete processes," *Expert Systems*, P. Dufour and A. van Lamsmeerde, Eds., Academic Press, 1986.
[23] B. R. Mitchell, *European Historical Statistics, 1750-1975*. New York, NY: Facts on File, 1980.
[24] A. J. Morgan, "Predicting the behavior of dynamic systems with qualitative vectors," *Advances in Artificial Intelligence*, J. Hallam, and C. S. Mellish, Eds., Chichester, West Sussex: J. Wiley, 1987.
[25] D. B. Osteyee and I. J. Good, *Information, Weight of Evidence, the Singularity between Probability Measures and Signal Detection*, Berlin: Springer-Verlag, 1974.
[26] S. Persson, "Some sequence extrapolating programs: A Study of representation and modeling in inquiring systems", Rept. No. STAN-CS-66-050, Department of Computer Science, Stanford University, Stanford, CA, 1966.
[27] M. Pivar and M. Finkelstein, "Automation using LISP, of inductive inference on sequences," in *The programming language LISP*, E.C. Berkeley and D. Bobrow, Eds., Cambridge, Mass:Information International, 1964.
[28] I. R. Racer, and J. E. Gaffney Jr, "The Potential role of artificial intelligence/expert systems in the warning and forecast operations of the national weather service," in *Proc. IEEE Experts Systems in Government Symposium* (Washington, D.C.), pp. 578-586, 1985.
[29] H. T. Reynolds, *Analysis of Nominal Data*. Beverly Hills, CA: Sage, 1984.
[30] H. A. Simon and K. Kotovsky, "Human acquisition of concepts for sequential patterns," *Psychological Review*, vol. 70, pp. 534-546, 1963.
[31] J. H. Slater, "Qualitative physics and the prediction of structural behavior," in *Expert Systems in Civil Engineering*, N. Kostem and M.L. Maher, Eds., New York, NY: ASCE, pp. 239-248, 1986.
[32] G. F. Swetnam and E. F. Dombroski, "An Expert System to Support the Forecasting of Upslope Snow Storms," in *Proceedings IEEE Experts Systems in Government Symposium* (Washington, D.C.), pp. 567-572, 1985.
[33] S. A. Vere, "Induction of relational productions in the presence of background information," *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, 1975.
[34] D. C. C. Wang and A. K. C. Wong, "Classification of discrete data with feature space transformation", *IEEE Transactions on Automatic Control*, vol. 24, no. 3, 1979.
[35] D. S. Williams, "Computer program organization induced from problem examples," in *Representation and Meaning: Experiments with Information Processing Systems*, H.A. Simon and L. Siklossy, Eds., NJ: Englewood Cliffs, pp. 143-205, 1972.
[36] G. Wright, P. Ayton and P. Whalley, "A General purpose computer aid to judgemental forecasting: rationale and procedures," *Decision Support Systems*, vol. 1, no. 4, 1985.
[37] A. K. C. Wong and D. K. Y. Chiu, "Synthesizing statistical knowledge from incomplete mixed-mode data," in *IEEE Transactions on Patt. Anal. and Mach. Intell.*, vol. 9, no. 6, pp. 796-805, 1987.
[38] S. Zubrick, "Validation of a weather forefasting expert system," In *Machine Intelligence 11*, American Elsvier, pp. 391-422, 1988.

Keith C. C. Chan received the B.Math (Hons.) degree in Computer Science and Statistics from the University of Waterloo, Waterloo, Ontario, Canada, in 1984. He received the M.A.Sc. and Ph.D. degrees from the same university in 1985 and 1989, respectively, in Systems Design Engineering. He was a research assistant in the Pattern Analysis and Machine Intelligence Laboratory of the Institute for Computer Research at the University of Waterloo from 1984 to 1989.

Following graduation, Dr. Chan joined the IBM Canada Laboratory, where he was involved in software development projects in the Image Systems Center and the Application Development Technology Center. He joined the Department of Electrical and Computer Engineering, Ryerson Polytechnic University, Toronto, Ontario, Canada in 1993 and is currently an Associate Professor in the department. He has been an adjunct faculty in the Department of Systems Design Engineering, University of Waterloo, since 1991 and he is also currently an adjunct faculty in the Department of Electrical Engineering at the University of Western Ontario, Ontario, Canada. His research interests include machine learning, neural networks, fuzzy systems, pattern recognition and computer vision.

Andrew K. C. Wong (M'79) received his Ph.D. from Carnegie Mellon University, Pittsburgh, PA, in 1968, and taught therte for several years thereafter. He is currently a professor of Systems Design Engineering and the director of the Pattern Analysis and Machine Intelligence Laboratory at the University of Waterloo and an honorable professor at the University of Hull, UK. Dr. Wong has authored and coauthored chapters and sections in a number of books on engineering and computer science and has published many articles in scientific journals and conference proceedings. He is the 1991 recipient of the Federation of Chinese Canadian Professionals Award of Merit.

David K. Y. Chiu (M'88) received his M.Sc. degree in computer science from Queen's University and his Ph.D. degree in engineering from the University of Waterloo, Ontario, Canada. He is currently an associate professor in the Department of Computing and Information Science at the University of Guelph, Guelph, Ontatio, Canada. He is associted with the PAMI Laboratory at the University of Waterloo and was a visitng researcher to the Electrotechnical Laboratory in Japan in 1992 under a Science and Technology Agency (STA) Fellowship program. His research interests are in the general area of pattern analysis, learning and knowledge discovery, and their applications to image analysis and computational molecular biology.