Abstract: Given an ordered sequence of objects (numbers, letters, symbols, observations, events, etc.) which is described by one or more attributes and is generated either deterministically or probabilistically. Suppose that the characteristics of the objects in the sequence is determined, to a certain extent, by the previous objects, the purpose of this paper is to present an inductive learning algorithm to analyze the inherent patterns in the sequence and to predict future objects based on these patterns. This inductive learning algorithm is divided into three phases: 1) detection of underlying patterns in a sequence of objects; 2) construction of rules, based on the detected patterns, that describes the generation process of the sequence; and 3) use of these rules to predict the characteristics of the future objects. The learning algorithm has been implemented in a program known as the OBSERVER and it has been tested with both simulated and real-life data. The experimental results show that the OBSERVER is capable of discovering hidden patterns and explaining the behavior of certain sequence generating processes that a user is not immediately aware of or fully understood. For this reason, the OBSERVER may be employed to solve complex real world problems where predictions have to be made in the presence of uncertainty. It can also be employed, like some inductive learning programs, to be an aid for the knowledge acquisition process in the construction of knowledge-based systems.
1. Introduction And Background

An inductive learning system is usually considered to have attained a concept if it is able to distinguish objects that embody the concept from those that do not. However, Simon and Kotovsky pointed out that there are certain kinds of concepts which we commonly measure attainment by the ability of a system to produce an object satisfying the concept rather than the mere ability to identify an object as belonging to that concept [14]. For example, a system is considered to have acquired the concept 'simple alternation of a and b' embodied in the sequence 'ababababa-' only when it is able to extrapolate the sequence by producing the next characters (i.e. ba) that satisfy the concept.

Since the ability of a learning system to acquire the concept underlying an ordered sequence of objects permits it to predict future objects based on the acquired knowledge, the task that such a system performs may be referred to as prediction. We have identified two different types of prediction problems: the deterministic prediction (DP) problem and the probabilistic prediction (PP) problem. Predicting succeeding letters in a letter sequence such as "ababababa-" is deterministic since they are completely determined by the preceding ones. Predicting weather conditions based on past records is, however, probabilistic since weather forecasts cannot usually be made with complete certainty. In this paper, an inductive learning algorithm that is able to handle these problems is presented.

2. Survey of Related Approaches

The DP problem was first considered by some psychologists with the intention to model human cognitive processes. For example, Simon and Kotovsky [8,14] proposed, and implemented in Concept Former, a model of human behavior in acquiring concepts from letter sequences. The Concept Former is then able to extrapolate certain letter sequences based on the periodic patterns detected in them. In [4,5,7,12,13,15,18], other strategies for solving the DP problem involving number or letter sequences are described.

To deal with DP problems other than number- or letter-series extrapolation, inductive learning programs such as the THOTH [16] can be employed. THOTH is capable of finding the maximal conjunctive generalizations between pairs of objects in a sequence so that, based on these generalizations, the characteristics of future objects may be predicted. The limitation of the THOTH is that it can only handle prediction problems in which all the characteristics of an object in a sequence can be determined exactly, and with complete certainty, by the preceding ones.

To deal with uncertainty, the SPARC program has been developed [3,9]. It is capable of handling a subset of the PP problem in which a sequence of objects are described by two different types of attributes: 1) those whose future values are completely determinable by the attributes of the preceding objects and 2) those whose future values are completely non-deterministic. In order to make predictions, SPARC employs a model-directed approach to guide the search for suitable sequence-generating rules in a predetermined search space consisting of all possible such rules. Since the search space may be very large (10^137 in the application described in [3]), its size is reduced by assuming that the possible sequence-generating rules are of three different types: periodic, decomposition and conjunctive normal form rules. Clearly, in problem domains where such an assumption cannot be validly made, the SPARC would be unable to discover the correct rules [3]. This is particularly true if the objects in a sequence are generated probabilistically. In the following, we will describe our approach to handling the PP problem of a less restricted type than that handled by the SPARC. That is, besides being completely deterministic or completely non-deterministic, the attributes of an object is allowed to depend on that of the preceding ones only to a certain degree (i.e. probabilistically).

3. Description of Object Sequence

To illustrate the proposed learning algorithm, suppose that there is an ordered sequence S of N objects (letters, numbers, observations, events, etc.), \( obj_1, \ldots, obj_p, \ldots, obj_N \), where \( obj_p \) is located at position \( p \) in S. Suppose also that these objects are described by \( n \) distinct attributes, \( Attr_1, \ldots, Attr_n \), and that in any instantiation of object description, an attribute \( Attr_j \) takes on a specific value, \( val_j \in \text{domain}(Attr_j) = \{ k \mid k = 1, \ldots, J \} \), which may be numerical or symbolical, or both. If we further assume that the \( N \) objects are generated in such a way that the characteristics, \( Attr_j \), \( j = 1, \ldots, n \) of an object, \( obj_p \) at a certain position \( p \) depends, probabilistically, on that of \( obj_1, \ldots, obj_{p-1}, obj_p \) preceding it, then the PP problem we are concerned with is to find a set of rules that describes how \( S \) sequence is generated so that, based on these rules, the characteristics of the future object can be accurately predicted.

4. The Inductive Learning Algorithm

The proposed learning algorithm for solving the PP problem can be divided into three phases: 1) detection of underlying patterns in a sequences of objects; 2) construction of rules that describes the generation process of the sequence based on the detected patterns; and 3) use of these rules to predict the characteristics of the future objects.

4.1. Detection of Patterns

To detect the patterns underlying a sequence of objects, we have to know how the attributes of an object are dependent on those before it. Obviously, if the \( i \)th attribute of an object which takes on the value \( v_i \) is always preceded at \( r \) positions in front by an object whose \( j \)th attribute takes on the value \( v_j \), we can conclude that \( v_i \) is dependent on \( v_j \). We can also say that such a dependence relation is perfect and positive. If, however, it is observed that \( v_i \) is never followed by \( v_j \) at \( r \) positions later, then we can also conclude that \( v_i \) is independent on \( v_j \). This is because whenever an object is observed to have the characteristic \( v_i \), the object that is located at \( r \) positions later in the sequence is not expected to possess the attribute value \( v_i \). In this case, we say that such a dependence relation is perfect and negative.

For the PP problem, however, dependence relations are rarely perfect even though they may either be positive or negative. In the presence of uncertainty, imperfect yet genuine dependence relationship between the attributes of an object and those of the objects preceding it have to be considered. Hence, if the number of objects in the sequence that are characterized by \( v_i \) and are preceded by objects with the characteristic \( v_j \) is significantly different from that expected of a hypothetical sequence generated under the assumption that the values are independent, we can conclude that, \( v_i \) is genuinely dependent on \( v_j \) at \( r \) positions in front.

To decide, objectively, if \( Attr_{(p-r)} \) of an object is dependent on \( Attr_p \) of the one at \( r \) positions in front, the chi-square test can be employed. Let \( o_{p} \) be the total number of objects in the sequence whose \( i \)th attribute takes on the
value, \( v_{ik} \), and are preceded, at \( r \) positions in front, by objects that has the characteristic, \( v_{jk} \). Let \( e_k \) be the total number of such objects that are expected under the assumption that \( \text{Attr}(p_{i+}) \) and \( \text{Attr}(p) \) are independent. Then, the goodness-of-fit chi square statistic can be computed as follows:

\[
X^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(o_{ik} - e_{ik})^2}{e_{ik}} = \sum_{i=1}^{k} \sum_{j=1}^{l} o_{ik}^2 - N \tag{1}
\]

where \( N = \sum_{i=1}^{k} \sum_{j=1}^{l} o_{ij} \). Whether or not the difference between what is observed and what is expected could have arisen by chance can be determined by comparing \( X^2 \) with the critical chi square value, \( \chi^2_{0.05} \). A value greater than \( \chi^2_{0.05} \) is significant as the observed value, \( v_{ik} \), differs from the expected value, \( e_{ik} \). If \( X^2 \) is greater than the critical value, one may conclude that \( \text{Attr}(p_{i+}) \) is genuinely dependent on \( \text{Attr}(p) \), otherwise, there is not enough evidence to support such a conclusion.

It should be noted, however, that a significant overall chi square test provides no information as to how the observed values of the \( i \)th attribute of an object is dependent on that of the \( j \)th attribute of the object at \( r \) positions in front. In other words, we may know that the \( j \)th attribute, as a whole, is important in determining the value of the \( i \)th attribute of the object at \( r \) positions later in the sequence, but we cannot tell whether or not a particular value, \( v_{jk} \), of it is important. This is especially true when both the \( i \)th and the \( j \)th attributes take on a large number of different values. In view of this, we propose a criterion for evaluating if a specific value of the \( i \)th attribute, \( v_{ik} \), of an object is genuinely dependent on the value, \( v_{jk} \), of the \( j \)th attribute of the object at \( r \) positions in front.

If the number of objects having the characteristic \( v_{ik} \) and is preceded, at \( r \) positions in front, by objects having the characteristic \( v_{jk} \) (i.e. \( o_{ik} \)), deviates significantly from the independence model (i.e. \( e_{ik} \)), then we can conclude that the dependence relation between them is genuine. Unfortunately, by looking only at the raw residuals [20], \( |o_{ik} - e_{ik}| \), we are not provided with any information on the relative degree of the discrepancy. In order to avoid being affected by the marginal totals, therefore, it is necessary to standardize the raw residuals in some ways. Haberman [6] recommended the adjusted standardized residual:

\[
d_k = \sqrt{\frac{v_k}{\sqrt{N}} \frac{1}{N} \frac{1}{\sqrt{1 - \frac{\lambda_i^2}{\lambda_k^2} + \frac{\lambda_i^2}{\lambda_k^2} - \frac{\lambda_i^2}{\lambda_k^2}}}} \tag{2}
\]

where \( v_k \) is the maximum likelihood estimate of the variance of \( v_k \), \( o_{ik} \) is the total number of objects in the sequence that possess the characteristic \( v_{ik} \) and \( \lambda_{ik} \) is the total number of objects in the sequence that have the characteristic \( v_{ik} \). Since \( d_k \) has an approximate normal distribution with mean 0 and variance 1, one could be able to determine which specific values of the \( i \)th attribute of an object and that of the \( j \)th attribute of another object at \( r \) positions in front are dependent on each other by searching for residuals that deviate from the model of independence at a significance level of 0.05. A value greater than 1.96 (the 95 percentile of the standard normal distribution) not only suggests that there is a significant discrepancy between the observed and the expected number of objects having the value \( v_{ik} \) and are preceded, at \( r \) positions in front, by objects characterized by \( v_{jk} \), it also indicates that the dependence relationship between them is positive and genuine. In other words, relative to the independence model, we may expect a higher proportion of objects in the sequence that have the characteristic \( v_{ik} \) to be followed, at \( r \) positions later, by objects whose \( i \)th attribute takes on the characteristic \( v_{jk} \). Similarly, a value smaller than -1.96 tells us that \( v_{ik} \) is negatively dependent on \( v_{jk} \). Therefore, we can expect a relatively high proportion of the objects that are preceded, at \( r \) positions in front, by objects that possess the attribute value \( v_{jk} \) to lack the characteristic \( v_{ik} \). Since the presence or absence of a certain attribute value, \( v_{jk} \), in an object provides important information about what the attribute values of the object at \( r \) positions later in the sequence may take on, the value, \( v_{jk} \), is relevant for the learning process. It provides positive or negative evidence supporting or refuting the object that is characterized by it to be followed, at \( r \) positions later, by another object whose \( i \)th attribute takes on \( v_{ik} \).

4.2. Rule Generation Based On Detected Pattern

Since relevant attribute values are important in determining the characteristics of the objects at a later position in a sequence, a simple way to ensure that they are utilized in the prediction process is to represent each detected dependence relation by a rule of the following form:

If \(<\text{condition}> \text{then} <\text{conclusion}>\) with uncertainty \( W \).

where the condition part of the rule specifies the characteristic(s) which an object possesses so that the object at a certain position in the sequence will take on the attribute value predicted by the conclusion part of the rule. As an illustration, suppose that the attribute value, \( v_{jk} \), is found to be dependent on \( v_{ik} \), then this relationship can be represented as:

If \( \text{Attr}(p_{i+}) \) of an object is \( v_{ik} \), then \( \text{Attr}(p_{i+}) \) of an object located at \( r \) positions later in the sequence has the value \( v_{jk} \) with uncertainty \( W \).

\( W = W(\text{Attr}(p_{i+}) = v_{ik}/\text{Attr}(p_{i+}) \neq v_{jk}) \), measures the amount of positive or negative evidence provided by \( v_{jk} \) to support or refute the object at \( r \) positions later to have the characteristic, \( v_{ik} \). The derivation of it is based on an information theoretic measure known as the mutual information. The mutual information between \( v_{ik} \) and \( v_{jk} \) is defined as [11, 17]:

\[
I(\text{Attr}(p_{i+}) = v_{ik}; \text{Attr}(p_{i+}) = v_{jk}) = \log \frac{Pr(\text{Attr}(p_{i+}) = v_{ik}|\text{Attr}(p_{i+}) = v_{jk})}{Pr(\text{Attr}(p_{i+}) = v_{jk})}. \tag{3}
\]

It has a positive value if and only if \( Pr(\text{Attr}(p_{i+}) = v_{ik}|\text{Attr}(p_{i+}) = v_{jk}) > Pr(\text{Attr}(p_{i+}) = v_{jk}) \) otherwise it is either negative or has a value 0.\( I(\text{Attr}(p_{i+}) = v_{ik}; \text{Attr}(p_{i+}) = v_{jk}) \), intuitively, the decrease (if positive) or increase (if negative) in uncertainty about the prediction of an attribute of an object to take on a certain value given that the object at \( r \) positions in front possesses the characteristic, \( v_{ik} \). The mutual information measure can also be thought of as a measure of the decrease or the increase in the gain in information when an object takes on a certain value given that the object at \( r \) positions in front has the characteristic \( v_{jk} \). Based on the mutual information measure, the weight of evidence for or against a certain prediction for the attribute values of future objects can
be assessed.

Suppose \( v_i \) is genuinely dependent on \( v_{ij} \), then a quantitative measure of evidence, the weight of evidence provided by \( v_i \) in favor of the \( i \)th attribute of the object at \( r \) positions later in the sequence to take on the value \( v_i \) as opposed to its taking on some other values can be defined as follows [11]:

\[
W(Atrrj_{(g+r)}) = \frac{v_i / Atrrj_{(g+r)} - v_i}{Atrrj_{(g+r)} - v_i} - I(Atrrj_{(g+r)} - v_i; Atrrj_{(g+r)} - v_i).
\]

(4)

In other words, the weight of evidence may be interpreted as a measure of the difference in the gain of information when the \( i \)th attribute of an object takes on the value \( v_i \) and when it takes on other values given that the object that is at \( r \) positions in front has the characteristic \( v_{ij} \). The weight of evidence is positive if \( v_{ij} \) provides positive evidence supporting the \( i \)th attribute of the object at \( r \) positions later in the sequence to take on the value \( v_i \), otherwise, it is negative.

The prediction rules that are constructed by the above procedure describe the generating process of the sequence of objects probabilistically.

4.3. Prediction of Future Objects

Given a set of rules constructed based on the detected dependence 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 \( N \) objects. Suppose also that we are to predict the value of the \( i \)th attribute, \( Attr_{(n+h)} \), of an object, \( obj_{(n+h)} \), \( h \) positions after the most recent one, \( obj_{(n)} \). The prediction process begins by searching through the space of prediction rules to determine how the characteristics of \( obj_{(n)}, obj_{(n-1)}, \ldots, obj_{(n-(L-1))} \), where \( L \) is the maximum lookback specified by the user, may affect the value of the \( i \)th attribute of \( obj_{(n+h)} \).

The search process proceeds by matching the attribute values of \( obj_{(n+h)}, j=1, \ldots, p=q=N, N-1, \ldots, (N-L)+1 \), of the objects, \( obj_{(n)}, obj_{(n-1)}, \ldots, obj_{(n-(L-1))} \), against the subset of prediction rules whose conclusion parts predict what values the \( i \)th attribute of an object at \( h, h+1, \ldots, (h+L)-1 \) positions later will take on based on the values of the \( j \)th attribute stated in the condition parts. An attribute value that satisfies the condition part of a rule in such a subset, therefore, affects the value of the \( i \)th attribute of the object at \( n+h \). Hence, this value provides a certain amount of evidence, as is reflected by the weight of the rule, supporting or refuting the \( i \)th attribute to take on the value predicted by the conclusion part of the rule.

Since the PP problem is being dealt with, it is possible for the \( i \)th attribute to be predicted to take on several different values based on different attributes of different objects in \( S \). For this reason, there is a need to decide which specific value \( Attr_{(n+h)} \) is most likely to have. To do this, it is noted that the attribute values of the objects \( obj_{(n-1)}, \ldots, obj_{(n-(L-1))} \) that match the prediction rules can be considered as providing some evidence for or against \( Attr_{(n+h)} \) to take on a certain value. Suppose that only \( m \) of the all these objects, \( val_{j1}, \ldots, val_{j\bar{m}} \), \( \ldots, val_{(n)}, val_{j1} \langle val_{j2} \rangle, j=1, \ldots, n; p=q=N, N-1, \ldots, (N-L)+1 \), are found to match one or more prediction rules, then the decision as to which specific one it should take on can be made by combining and weighing these pieces of evidence.

To quantitatively estimate and combine these evidences 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 the strength of the positive evidences supporting a specific value of \( Attr_{(n+h)} \) and decreases with the number and the strength of the negative evidences refuting such a value. This measure, known also as the weight of evidence, is a modification from what is described in the last section [1,2]. It measures the evidences, provided by the attribute values, \( val_{jl}, j=1, \ldots, n; p=q=N, N-1, \ldots, (N-L)+1 \), of the \( (N-L)+1 \) of the objects \( obj_{j1}, obj_{j2}, \ldots, obj_{(N-L)+1} \) in favor of \( Attr_{(n+h)} \) taking on a certain value. It is defined as:

\[
W(Attr_{(n+h)}=val_{j1} | Attr_{(n+h)}=val_{jl}) = W(val_{j1} | val_{jl}) = W(val_{j1} | val_{jl}) \ldots W(val_{j1} | val_{j\bar{m}})
\]

(5)

If 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 \( obj_{j1}, obj_{j2}, \ldots, obj_{(N-L)+1} \), in favor of \( Attr_{(n+h)} \) taking the value \( val_{j1} \) 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 \( obj_{j1}, obj_{j2}, \ldots, obj_{(N-L)+1} \) that is relevant for the classification task. For this reason, we can write [1,2]:

\[
W(val_{j1} | val_{jl}) = \sum_{j=1}^{m} W(val_{j1} | val_{jl})
\]

(6)

Hence, intuitively, the total amount of evidence supporting or refuting \( Attr_{(n+h)} \) is equal to the sum of the individual pieces of evidence provided by each relevant attribute value of the objects, \( obj_{j1}, obj_{j2}, \ldots, obj_{(N-L)+1} \).

In brief, the strategy for predicting a certain characteristic of a future object based on a sequence of \( N \) objects can be summarized as follows. Given an object \( obj_{(n+h)} \) whose \( i \)th attribute is to be predicted based on the attribute of the \( L \) most recent objects, \( obj_{(n)}, obj_{(n-1)}, \ldots, obj_{(n-(L-1))} \), the set of prediction rules is searched to determine which characteristics of these \( L \) objects affect the value of \( Attr_{(n+h)} \) of \( obj_{(n+h)} \). 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 \( i \)th attribute at \( h, h+1, \ldots, (h+L)-1 \) positions later. Therefore, if an attribute value of one of the \( L \) objects satisfied the condition part of a rule, then it affects the value of the \( i \)th attribute of the object, \( obj_{(n+h)} \), at a later position specified by the rule. The attribute value can be considered as providing some evidence for or against \( Attr_{(n+h)} \) taking on the value predicted by the conclusion part of the rule. The strength of this evidence is given by the weight of the rules. Since the attributes of the objects, \( obj_{(n)}, obj_{(n-1)}, \ldots, obj_{(n-(L-1))} \), may or may not provide evidence, and even for those that do, they may support different values, these evidences are quantitatively measured and combined for comparison in order to find the most probable value that \( Attr_{(n+h)} \) should take on. Based on the weight of evidence measure, \( Attr_{(n+h)} \) is predicted to take on the value \( v_{ij} \), if:

\[
W(Attr_{(n+h)}=v_{ij} | Attr_{(n+h)}=v_{(n+h)}) = W(val_{j1} | val_{jl}) > W(Attr_{(n+h)}=v_{ij} | Attr_{(n+h)}=v_{(n+h)}) = W(val_{j1} | val_{jl}) > W(Attr_{(n+h)}=v_{ij} | Attr_{(n+h)}=v_{(n+h)}) = W(val_{j1} | val_{jl})
\]

226
\[ q=1, 2, \cdots, I \text{ and } q \neq i \] (7)

where \( i' \) denotes the number of values of the \( i \)th attribute that are dependent on that of the \( L \) most recently object 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 \( Attr_{j}(n+k) \) may take on. Furthermore, if there is no evidence for or against any specific value of the \( i \)th attribute, prediction will be refrained instead of furnishing an inaccurate one. If it happens that there is no relevant value for determining the future value of the \( i \)th attribute of \( obj(n+k) \), then either the generation of the sequence is completely non-deterministic or there is insufficient generated objects for any prediction to be made.

5. Examples

The above inductive learning algorithm has been implemented in a program called the OBSERVER. It has been tested with both simulated and real-world data.

5.1. Example 1: A Sequence of Movements

Suppose four objects 1, 2, 3, and 4, are moving around in a bounded area and the OBSERVER is taking a snapshot of it at fixed time intervals (Figure 1). In order to demonstrate the capability of the proposed learning method to discover rules that govern the interrelations of the movements of a set of objects over a period of time, the PP problem is simulated. The data for this experiment is generated probabilistically in such a way that object 1 moves randomly. Object 2 moves towards the same direction as 1 all the time. Object 3 moves to the same direction as 2 about 55% of the time whenever the previous move of 2 is to E, S, or W, but however, if 2 moved to the N, then 3 may move to any direction except to the N. As for 4, it moves to a direction that is opposite to that of 3 whenever 3 moved to 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 \( r \) can be described by four attributes, \( Attr_{j}(r)=1, \ldots, 4 \) corresponding to the movements of the four objects so that domain(\( Attr_{j} \)) = \{E, S, W, N\}. To determine the future position of an object, say 3, we have to know how it 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, it is discovered that there is a significant dependence relationship between the movements of 3 and that of 2 at a time difference of one unit (\( X^2_{0.05} \)). However, basing 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 is described in the last section has to be conducted by analyzing the adjusted residuals as given in Table 1 so as 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 deviates, we observe that \( d_E, d_S, d_W, d_N \), 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, \( d_E \), conclusively state that whether 3 moves to \( E \), or not depends on whether the last movement of 2 is towards the \( E \), yet there are obvious reasons for such premises.

<table>
<thead>
<tr>
<th>( Attr_{j}(n+k) )</th>
<th>( d_E )</th>
<th>( d_S )</th>
<th>( d_W )</th>
<th>( d_N )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Attr_{3}(n+1) )</td>
<td>2.5</td>
<td>-1.2</td>
<td>-1.5</td>
<td>0.1</td>
</tr>
<tr>
<td>( Attr_{3}(n+1) )</td>
<td>-1.6</td>
<td>2.6</td>
<td>-1.7</td>
<td>0.5</td>
</tr>
<tr>
<td>( Attr_{3}(n+1) )</td>
<td>-1.6</td>
<td>-1.8</td>
<td>2.7</td>
<td>0.9</td>
</tr>
<tr>
<td>( Attr_{4}(n+1) )</td>
<td>0.7</td>
<td>0.5</td>
<td>1.0</td>
<td>-2.0</td>
</tr>
</tbody>
</table>

Also, from the signs of the deviates, we can conclude that the dependence of \( E, S, W \) of 3 on \( E, S, W \) of 2, respectively, are positive whereas \( N \) of 3 is negatively dependent on \( N \) of 2. This suggests that the movement of 2 to \( E, S, W \) provides evidence supporting that the next movement of 3 is towards \( E, S \) or \( W \) respectively whereas the movement of 2 to the \( N \) implies that there is negative evidence against the next movement of 3 to be \( N \). In a similar way, other genuine dependence relationship between the movements of different objects at different times are detected and the probabilistic patterns by which the data is generated is successfully discovered.

Based on the detected genuine 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 that the next movement of 3 is to the \( W \) is with uncertainty 1.66 where 1.66 is the weight of evidence for 3 to move to the \( W \) as opposed to its moving to \( E, S \), or \( N \), provided that the last movement of 2 is also towards the \( W \).

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 Figure 1.

To do this, we have to know which movements of 3 are dependent on the previous movements of other objects. In other words, the maximum look-back is 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 favor of 3 moving to a direction, say \( W \), we compute:

\[ W(Attr_{3}(n+1)=W) = W(Attr_{3}(n+1)\neq W | Attr_{2}(n)=E) \]

\[ W(Attr_{2}(n)=W) \]

\[ W(Attr_{3}(n+1)=W | Attr_{3}(n+1)\neq W | Attr_{2}(n)=W) = 1.66 \]

By a similar procedure, the total weight of evidence in favor of 3 moving to \( E, S \), and \( N \) as opposed to its moving to other directions can be determined and 3 is predicted to move to \( W \) which is a direction supported by the strongest evidences. The next movements of all the other objects are also correctly predicted with the exception of 1. No pattern can be detected concerning the movement of 1 as it moves independently of the others in a completely random manner.

It is interesting to note that the movements of the objects are correctly predicted even though the relative positions occupied by them in the 5x5 matrix have never been observed before.

5.2. Experiment 2: A Sequence of Geometric Figures

Suppose that the OBSERVER is given snapshots of an ongoing process which generates a sequence of geometric figures (Figure 2) [3]. Suppose also that each of the figure is described by four attributes: shape, number of nodes, node texture and orientation [3]. The problem is to predict the characteristics of the figure at \( p+1 \). Based on the learn-
ing algorithm described in the last section, the OBSERVER is able to discover the patterns underlying the sequence of geometric figures: the figures in the sequence can be grouped together into subsequences. For example, if the figures in each subsequence have textures in the order: solid black, blank, cross, and the corresponding shapes are always: T-junction, a shape other than T-junction or bar, 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, the figure at position $p+1$ is predicted to have 8 blank nodes and that its shape is neither a T-junction nor a bar (i.e. there is negative evidence against the future figure taking either of the two shapes).

Since the shape of the figure cannot be determined, its orientation with respect to its last appearance is also not predicted.

It must be noted that the maximum lookback is taken to be 1 in this experiment. However, all the predictions can still be accurately made. 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 the number of nodes of the figure at position $p-2$. In case of the OBSERVER, however, it 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 by the authors before the experiment.

The SPARC has also been tested with this experiment. However, it is unable to correctly predict the characteristics of the next figure. This is due to the weaknesses of the SPARC program that we discussed earlier in Section 2. The inability of the SPARC to form composite models results in its not being able to find sequence-generating rule in a sequence of objects which involves a nested periodic structures [3].

5.3. Experiment 3: A Sequence of Playing Cards

Suppose that the OBSERVER is given a layout (the mainline) of a game of Eleusis (Figure 3) [3]. Suppose also that each card of the sequence is described by the simple attributes: suit, rank, color, facedness, parity, primeness and the compound 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 their ranks. The problem is to find the secret rule that governs the layout of the cards and to predict the next figure in the sequence.

Based on the learning algorithm described in the last section, the OBSERVER is able to discover the secret rule which 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 suit higher (modulo 4). Based on these secret rule, 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’. The exact rank of the card is not predicted due to inadequate evidence.

The results of this experiment shows that the OBSERVER is able to discover rules which are not immediately obvious to a human. It also performs better than the SPARC program which is unable to find the secret rule for this game [3].

5.4. Experiment 4: 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 [10] 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 are described by a total of 12 attributes are available. Each of these attributes represents the mean July temperature of a certain year in a city in Europe. The task for the OBSERVER is to discover if there is any underlying pattern in the data and to forecast the future based on the detected patterns.

Since it is more meaningful, especially in long term forecasting, for mean temperature to be predicted to be in 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 a lot of 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 [19]. This procedure allows the temperature scale to be divided into different intervals while at the same time reduces the loss of information to minimum. By such a procedure, the recorded mean temperature for each city were grouped into four intervals before the experiment and each of these intervals represents a certain range of temperature as shown in Table 2.

<table>
<thead>
<tr>
<th>City</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copenhagen</td>
<td>12.1-14.0</td>
<td>14.1-16.0</td>
<td>16.1-17.2</td>
<td>17.3-17.9</td>
</tr>
<tr>
<td>Edinburgh</td>
<td>15.6-18.3</td>
<td>18.4-19.2</td>
<td>19.3-20.5</td>
<td>20.6-23.4</td>
</tr>
<tr>
<td>Geneva</td>
<td>12.9-16.0</td>
<td>16.1-17.0</td>
<td>17.1-18.4</td>
<td>18.5-21.4</td>
</tr>
<tr>
<td>Stockholm</td>
<td>14.7-15.3</td>
<td>15.4-18.4</td>
<td>18.5-20.5</td>
<td>20.6-23.4</td>
</tr>
<tr>
<td>London</td>
<td>20.5-23.6</td>
<td>23.7-24.4</td>
<td>24.5-25.5</td>
<td>25.3-27.6</td>
</tr>
<tr>
<td>Rome</td>
<td>19.8-21.8</td>
<td>21.9-22.6</td>
<td>22.7-23.3</td>
<td>23.4-25.6</td>
</tr>
<tr>
<td>Marseilles</td>
<td>18.9-22.9</td>
<td>23.0-23.7</td>
<td>23.8-25.0</td>
<td>25.1-28.3</td>
</tr>
<tr>
<td>Milan</td>
<td>15.2-17.4</td>
<td>17.5-18.3</td>
<td>18.4-19.7</td>
<td>19.8-22.0</td>
</tr>
<tr>
<td>Paris</td>
<td>5.4-7.6</td>
<td>7.7-8.9</td>
<td>9.0-9.9</td>
<td>10.0-13.6</td>
</tr>
<tr>
<td>Berlin</td>
<td>4.3-8.9</td>
<td>9.0-9.6</td>
<td>9.9-10.9</td>
<td>11.0-14.6</td>
</tr>
<tr>
<td>Vienna</td>
<td>3.10-6.0</td>
<td>6.1-6.9</td>
<td>7.0-7.9</td>
<td>8.0-12.7</td>
</tr>
<tr>
<td>Oslo</td>
<td>3.10-6.0</td>
<td>6.1-6.9</td>
<td>7.0-7.9</td>
<td>8.0-12.7</td>
</tr>
</tbody>
</table>

By applying the inductive learning algorithm described above, the OBSERVER is able to discover how the temperature of the different cities affects 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 applying 15% of the available weather records for testing. The mean temperature of each of the 12 cities for the last 27 years from 1949 to 1975 (15%) are predicted based on the weather records of the previous years (the maximum lookback is taken to be 6). The results for the tests are shown in Table 3. It should be noted that the predictions made by the OBSERVER is particularly accurate in forecasting the mean temperatures of the cities such as Stockholm, Rome, Milan, Berlin and Oslo. The percentages of accurate predictions are as high as 74%. It must also be noted that the mean temperatures for some cities such as Geneva, London and Marseilles are difficult to be predicted accurately. Their temperature seems to be independent of other cities and are probably influenced by other geographical factors which were not taken into consideration. In view of the difficulty in long-term weather forecasting, the experimental results are very encouraging.
vides first clues in the search for underlying causal relations that a human is unlikely to notice.

The proposed learning algorithm 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. It can also be employed, like many inductive learning programs, to be an aid for the knowledge acquisition process in the construction of knowledge-based systems. The detection of probabilistic relations also provides first clues in the search for underlying causal relations that a human is unlikely to notice.

Table 3. Results for experiment 4

<table>
<thead>
<tr>
<th>OBSERVER</th>
<th>correct</th>
<th>incorrect</th>
<th>reject</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copenhagen</td>
<td>28.9</td>
<td>66.7</td>
<td>3.7</td>
</tr>
<tr>
<td>Edinburgh</td>
<td>59.3</td>
<td>29.6</td>
<td>11.1</td>
</tr>
<tr>
<td>Geneva</td>
<td>18.5</td>
<td>77.8</td>
<td>3.7</td>
</tr>
<tr>
<td>Stockholm</td>
<td>81.5</td>
<td>14.8</td>
<td>3.7</td>
</tr>
<tr>
<td>London</td>
<td>22.2</td>
<td>70.4</td>
<td>7.4</td>
</tr>
<tr>
<td>Rome</td>
<td>81.5</td>
<td>14.8</td>
<td>3.7</td>
</tr>
<tr>
<td>Marseilles</td>
<td>18.5</td>
<td>63.0</td>
<td>18.5</td>
</tr>
<tr>
<td>Milan</td>
<td>88.9</td>
<td>11.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Paris</td>
<td>29.6</td>
<td>59.3</td>
<td>11.1</td>
</tr>
<tr>
<td>Berlin</td>
<td>74.1</td>
<td>22.2</td>
<td>3.7</td>
</tr>
<tr>
<td>Vienna</td>
<td>25.9</td>
<td>51.9</td>
<td>22.2</td>
</tr>
<tr>
<td>Oslo</td>
<td>74.1</td>
<td>14.8</td>
<td>11.1</td>
</tr>
</tbody>
</table>

6. Conclusion

A three-phase inductive learning algorithm that is able to handle the PP problem is described in this paper. It is capable of discovering the probabilistic patterns inherent in a sequence of objects and is able to construct, without supervision, prediction rules based on these patterns. The learning algorithm has been implemented in the program OBSERVER. It has been tested with both simulated and real-life data. The experimental results show that it is capable of explaining the behavior of certain sequence generating processes that a user is not immediately aware of or fully understood.

The proposed learning algorithm 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. It can also be employed, like many inductive learning programs, to be an aid for the knowledge acquisition process in the construction of knowledge-based systems. The detection of probabilistic relations also provides first clues in the search for underlying causal relations that a human is unlikely to notice.

References


