Probabilistic Finite Automata and Randomness in Nature: a New Approach in the Modelling and Prediction of Climatic Parameters[1]

L.Mora-López[a], **R.Morales-Bueno**[a], **M.Sidrach-de-Cardona**[b], **F.Triguero**[a]

[a] *Dpto.Lenguajes y C.Computación (llanos@lcc.uma.es, morales@lcc.uma.es, triguero@ctima.uma.es)*
[b] *Dpto. Física Aplicada II. (msidrach@ctima.uma.es)*
[a,b] *E.T.S.I.Informática.Universidad de Málaga. Campus Teatinos. 29071 Málaga*

Abstract: A model to characterize and predict continuous time series from machine learning techniques is proposed. This model includes the following three steps: dynamic discretization of continuous values, construction of probabilistic finite automata and prediction of new series with randomness. The first problem in most models from machine learning is that they are developed for discrete values; however, most phenomena in nature are continuous. To convert these continuous values into discrete values a dynamic discretization method has been used. With the obtained discrete series, we have built probabilistic finite automata which include all the representative information which the series contain. The use of probabilistic finite automata allows us to consider, in an easy way, the different relationships between the values in the series for different environmental conditions. The learning algorithm to build these automata is polynomial in the sample size. An algorithm to predict new series has been proposed. This algorithm incorporates the randomness in nature: values are generated using the cumulative probability distribution function -included in the automata- and a random number to select the new value. After finishing the three steps of the model, the similarity between the predicted series and the real ones has been checked. For this, a new adaptable test based on the classical Kolmogorov-Smirnov two-sample test has been developed; this test takes into account the continuous nature of climatic data. The cumulative distribution function of observed and generated series has been compared using the concept of indistinguishable values. Finally, the proposed model has been applied in a practical cases: the study of hourly global solar radiation series.

Keywords: Machine Learning, Modelling Climatic Data, Time Series

1. INTRODUCTION

The fundamental idea in this paper is the use of probabilistic finite automata (PFA) as a means of representing the relationships observed in climatic data series. PFAs are mathematical models used in the machine learning field.

Traditionally, the analysis of time series has been carried out using stochastic process theory. One of the most detailed analyses of statistical methods for time series research was done by Box et al [1976].

The goal of data analysis by time series is to find models which are able to reproduce the statistical characteristics of the series. Moreover, these models allow us to predict the next value of the series from its predecessors.

Probabilistic finite automata have been used to model several types of natural sequences. Examples of such applications are: universal data compression, Rissman [1983], analysis of biological sequences, for DNA and proteins, Krogh et al. [1993], analysis of natural language, for handwriting and speech, Nadas [1984], Rabiner

[1] This work has been partially supported by FACA Project number PB98-0937-C04-01 of the CICYT, Spain. FACA is a part of the FRESCO project

[1994] and Ron et al. [1998], etc. Different classes of automata have been developed. For instance, acyclic probabilistic finite automata have been used for modeling distributions on short sequences, Ron et al. [1998]; probabilistic suffix automata, based on variable order Markov models, have been used to construct a model of the English language, Ron et al. [1994]. All these automata allow us to take into account the temporal relationships in a series.

These machine learning models are very useful to study systems in which the concept to learn presents probabilistic behaviour. The prediction of climatic variables is an example of these types of concepts. In these systems the recorded variables are insufficient to exactly determine the future values, due to the random nature of these variables. The systems in which these models can be used must have the following properties:

- Present probabilistic behaviour or uncertainty. This uncertainty can be due to several factors. For example, for the prediction of climatic variables the number of parameters which affect them is very high.
- Although there is uncertainty in these systems, there is always some structure within this uncertainty.

This paper describes how to use certain models from the machine learning field in the analysis and prediction of climatic parameters. The model we propose is based on the Probabilistic Finite Automata Theory. Our goal is to use PFAs to represent all the relationships observed in natural time series and to use these PFAs to predict new values of the series. Moreover, an adaptable test based on the classic Kolmogorov-Smirnov two-sample test has been used to check the proposed model. Finally, preliminary results of the model obtained for a climatic parameter are shown.

2. PROBABILISTIC FINITE AUTOMATA

2. 1 Introduction

We propose using a mathematical model called probabilistic finite automata (PFA). We propose the use of this mathematical model to represent a univariate time series. Formally, a PFA is a 5-tuple $(\Sigma,Q,\tau,\gamma,q_0)$ where:

- Σ is a finite alphabet; that is, a set of discrete symbols corresponding to the different continuous values of the analyzed parameter. The different symbols of Σ will be represented by x_i. For a series, the values observed can be $x_5x_3,...x_3$ To represent the different observable series for a period t_1 to t_m we will use the symbols $y_1y_2...y_m$. So, in the series $x_5x_3...x_3$, the symbol $y_1$corresponds to the value x_5, the symbol y_2 to x_3 and so on.
- Q is a finite collection of states. Each state corresponds to a subsequence of the discretized time series. The maximum size of a state -number of symbols- is bounded by a value N fixed in advance. This value is related to the number of previous values which will be considered to determine the next value in the series and depends on "memory" of the series.
- τ: $Q \times \Sigma \rightarrow Q$ is the transition function
- γ: $Q \times \Sigma \rightarrow [0,1]$ is the next symbol probability function
- $q_0 \in Q$, is the initial state

The function γ satisfies the following requirement: For every $q \in Q$ and for every $x_i \in \Sigma$, $\Sigma_{xi\in\Sigma}\ \gamma(q, x_i)=1$. Moreover, the following conditions are required:

- The transition function τ can be undefined only on states $q \in Q$ and symbols $x \in \Sigma$, for which $\gamma(q,x)=0$;

- The function τ can be extended to be defined on $Q \times \Sigma^*$ in the following recursive manner:

$$\tau(q,y_1,y_2,...y_t)= \tau(\tau(q,y_1,y_2,...,y_{t-1}),y_t).$$

where $yi \in \Sigma$. Graphically, each state is represented by a node and the edges going out of each state are labeled by symbols drawn from the alphabet. Moreover, each state has an associated probability vector which is composed of the probability of the next symbol for each of the symbols of the alphabet. For instance, in figure 1 a simple PFA is shown.

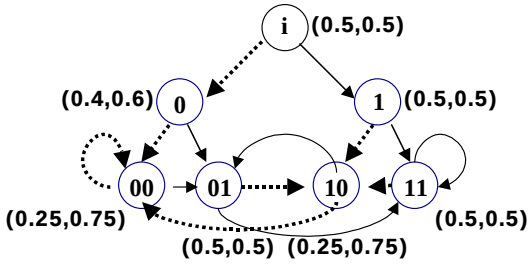

Figure 1. Example of probabilistic finite automata

In this PFA, the alphabet, Σ, is composed of the symbols 0 and 1. The states of the system, Q, are described in each node of the automata: initial (i), 0, 1, 00, 01, 10 and 11. For instance, the state labeled 01 corresponds to the following sequence of values in the series: 1 as the last value and 0 as the previous. The associated vectors at each state (node) are the probabilities which each symbol of the alphabet has to appear in the next moment, after the sequence of symbol that label the node has appeared. For instance, the node labeled with

10, has the associated vector (0.25,0.75); this means that if the current state is 10, then the next symbol can be 0, with a probability of 0.25 and 1 with a probability of 0.75. The continuous and discontinuous arrows represent the transition function between states (discontinuous for 0, and continuous for 1). For instance, if the current state is 10, and the next symbol is 0, then the following state will be labeled with 00; but if the next symbol is 1, then the following state will be labeled with 01.

In the PFA shown in Figure 1, the states 01 and 11 have the same probability vector as state 1. That is, when the symbol 1 appears, it is not necessary to know the preceding value to determine the probabilities of the next symbol, since in both cases, (0 or 1), the probabilities vector of the next symbol is (0.5,0.5). Therefore, the PFA of Figure 1 can be converted into the PFA shown in Figure 2.

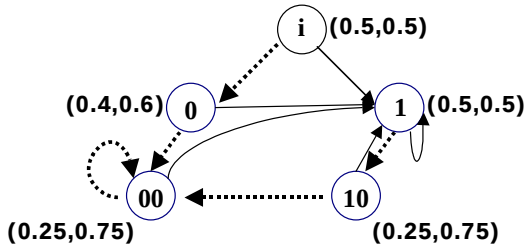

Figure 2. Simplified probabilistic finite automata

This class of PFA is used to represent variable order Markov models. These simplified automata are the automata proposed in this paper. They capture the same information with fewer states than the original automata. Moreover, they allow us to take into account, for each state, a different number of previous values in the series.

Let us define some concepts that we will use to build the PFAs for climatic data series. Let $\Sigma=\{x_1,x_2,...,x_n\}$ be the set of discrete values of the analized variable and Σ^* denotes the set of all possible sequences which can be obtained with these values. For any integer N, Σ^N denotes the set of all possible sequences of length N and $\Sigma^{\leq N}$ is the set of all possible sequences with length less than or equal to N. For any subsequence, Y, represented by $y_1...y_m$, where $y_i \in \Sigma$, the following notations will be used:

- The longest final subsequence of Y, different from Y, will be $final(Y)=y_2...y_m$
- The set of all final subsequences of Y will be, $last(Y)=\{y_i...y_m | 1 \leq i \leq m\}$

In the next section we explain how to build a PFA for a time series.

3. BUILDING PROBABILISTIC FINITE AUTOMATA

3.1 Algorithm to build probabilistic finite automata.

The following algorithm is used to construct the PFA:

1. Compute the series of discrete values.

2. Initialize the PFA with a node, with label null sequence.

3. The set *PSS* -Possible Subsequence Set- is initialized with all sequences of order 1. Each element in this set corresponds to a sequence of discrete values. Take *o=1* as the initial value of the order –that is, size of subsequences to consider.

4. If there are elements of order *o* in *PSS*, pick any of these elements, *Y*. Using all discrete sequences in the series, compute the frequency of *Y*. If 4.a and 4.b are true, then go to 5, else go to 6.

4.a The frequency of this sequence is greater than the threshold frequency.

4.b For some $x_p \in \Sigma$, the probability of occurrence of the subsequence Yx_p is *not equal* to the probability of the subsequence $final(Y)x_p$, s, that is:

$$P(x_p|Y) \neq P(x_p|final(Y)).$$

(*not equal*: when the ratio between the probabilities is significantly greater than one)

5.Do

5.a Add to the PFA a node, labeled with *Y*, and compute its corresponding probabilities vector.

5.b For each amplified sequence, Yx_p: if the probability of this augmented sequence is greater than the threshold probability, then include it in *PSS*.

6. Remove the analyzed subsequence, *Y*, from *PSS*.

7. If there are no elements of order *o* in *PSS*, add 1 to the value of *o*. If $o \leq N$ and there are elements of length *o* in *PSS*, then go to 4, else Stop.

3.2 Predicting new values

A PFA can be used as a mechanism for generating finite sequences of values in the following manner. Start from an initial value selected from the alphabet, called the initial state. If q_t is the current state, labeled by the sequence Y=y1...yt, then the next symbol is chosen (probabilistically) according to $\gamma(q_t,\cdot)$. If $x \in \Sigma$ is the chosen symbol, then the next state, q_{t+1}, is $\tau(q_t,x)$. The label of this new

state, Y', will be the longest final subsequence of Yx in the PFA, that is:

$$Y'=Max\{last(Yx)\}\in PFA.$$

The process continues until the length of the required sequence is reached.

Moreover, if $P^t(Y)$ denotes the probability that a PFA generates a sequence $Y=y_1...y_{t-1}y_t$, then:

$$P^t(Y)=\prod_{i=0}^{t-1}\gamma(q_i,y_{i+1}).$$

This definition implies that $P^t(\cdot)$ is in fact a probability distribution over the symbols of sequence, i.e.:

$$\sum_{Y\in\Sigma^*}P^t(Y)=1.$$

4. HOW THE MODEL CAN BE VALIDATED

For a recorded time series, the following steps must be followed to use the proposed model. First, if the time series has continuous values, then these values must be discretized. After this, the PFA is built using the discrete series. With the PFA and the generation method described above, new values for the time series can be generated.

In order to compare the simulated series to the real ones, several statistical tests can be used. The hypothesis that both series have the same mean and variance will be checked.

The frequency histograms of the recorded and simulated series are also analyzed. To make this comparison, we propose the use of an adaptable goodness-of-fit test, which is based on the two-sample Kolmogorov-Smirnov test, described in Rohatgi [1976]. The objective of this adaptable test is to determine if two distribution functions $F_Y(.)$ and $F_Z(.)$ are the same, except for possible changes in location and scale. Specifically, we have checked the null hypothesis that there exist two unknow values μ and σ such that Z_i and $\mu+\sigma Y_j$ have the same distribution. Using distribution functions it is possible to express our null hypotheses as follows:

$$H_0:\exists\mu\in\Re \text{ and } \sigma\in(0,+\infty)/\forall u\in\Re$$

$$F_X(u)=F_Y\left(\frac{u-\mu}{\sigma}\right)$$

Replacing unknow parameters μ and σ by estimates introduces additional random terms in the statistic. Therefore, to obtain the critical values that must be used in the test, we propose using a bootstrap procedure.

5. A PRACTICAL CASE: USING PROBABILISTIC FINITE AUTOMATA FOR CLIMATIC DATA

The probabilistic finite automata presented in the previous sections have been used to characterize and predict a climatic variable; the hourly global radiation received on a surface on the ground. For this variable, time series are recorded by meteorological stations at regular time intervals.

We need a stationary time series. From the original series we have calculated the series of the hourly clearness index, which are stationary. The following question - which we have solved- is the discretization of these series. The recorded values are continuous whereas the proposed mathematical model uses discrete values. The discretization method used is explained later. The PFAs have been built using the discrete series obtained and new values of the series generated. Finally, we have checked these values using several tests.

5.1 Data set

The data of the hourly exposure series of global radiation, $\{G_h(t)\}$, which are used in this work were recorded over several years in nine Spanish meteorological stations. In total, 745 months were accounted for. The pertinent latitudes range from 36ºN to 44ºN. The annual average values of daily global solar radiation for these locations range from 11MJm^{-2} to 18 MJm^{-2}.

5.2 Discretization of the series

The goal is to use an effective and efficient method to transform continuous values into discrete ones using the overall information included in the series and, when possible, feedback with the learning system. To do this, the discrete value which corresponds to a continuous value has been calculated using qualitative reasoning, taking into account the evolution of the series. Qualitative models have been used in different areas in order to obtain a representation of the domain based on properties (qualities) of the systems; see, for instance, Forbus[1984], Keller et al. [1984] and Kuipers[1984].

We have used a qualitative dynamic discrete conversion method, described in Mora et al.[2000]. It is dynamic because the discrete value associated to a particular continuous value can change over

time: that is, the same continuous value can be discretized into different values, depending on the previous values observed in the series. It is qualitative because only those changes which are qualitatively significant appear in the discretized series.

5.3 PFA for global hourly solar radiation series

The parameter used to build the PFA is the hourly clearness index, defined as:

$$K_h = G_h / G_{h,0}$$

where G_h is the hourly global radiation and $G_{h,0}$ is the extraterrestrial hourly global radiation.

The alphabet of the PFA is:

$$\Sigma=\{0,1,...,7\}$$

The relationship between the values of the clearness index and the symbols of the alphabet is the following. For the first symbol of a series, the discrete value of the series will be calculated using:

$$Y_h = \begin{cases} 0 & 0 \le K_h < 0.35 \\ \left\lfloor \dfrac{K_h - 0.35}{0.05} \right\rfloor + 1 & 0.35 \le K_h < 0.65 \\ 7 & K_{h \ge 0.65} \end{cases}$$

where $\lfloor A \rfloor$ means the integer value of A. For the following values of the continuous series the discretes values will be obtained using the algorithm described in Mora et al.[2000] and the intervals aforementioned.

Using these expressions and the hourly clearness index series, the discrete series $\{Y_h\}$ are obtained. From all possible subsequences observed in the series, only those with a sufficient probability will be used to build the PFA. This threshold of probability must be defined when the PFA is built.

The monthly series of the hourly clearness index have been grouped using the monthly mean value of the hourly clearness index. The ranges for each group are the same as those defined for the discretization of this parameter. For every interval, one PFA has been built.

5.4 Predicting new values

To generate new series we need an initial state. The initial state is the discrete value corresponding to the mean value of the clearness index for each series. Let q_t be the current state. The next symbol, y_{t+1}, is generated as follows: first, a random number $r \in [0,1]$ is generated. Then, we chose the only component of probabilities vector -for the current state, q_t- which satisfies:

$$y_{t+1} = x_j \mid \sum_{i=1}^{j} \gamma(q_t, x_i) \ge r \ AND \ x_{j-1} \mid \sum_{i=1}^{j-1} \gamma(q_t, x_i) < r$$

5.5 Results

To select the best values of the parameters to build the PFAs, we have checked the results obtained with different values of these parameters. The values we have used are:

- Order of the PFA: from 2 to 14.
- Threshold -minimum number of appearances of a sequence- from 1 to 5.

For most of the intervals, if the order used for the PFA is 2, the results are similar to those when the order is 4; however, for intervals 5 and 6, using order 4, the PFA captures the relationship observed in the series better than using order 2. Thus, the selected order (maximum) for the PFA is 4. The selected minimum number of appearances -required to use a sequence to build a PFA- is 2.

With the built PFAs, new sequences of the hourly clearness index have been generated. The original and generated series have been compared using the statistical test described above.

The results obtained for each interval of the clearness index are shown in Table 1.

Table 1. Results obtained for each interval of the clearness index, with N=4 and threshold=2. The third column shows the number of months predicted as being similar to the real ones. The fourth column shows the percentage.

Interval	**Months**	**Similar**	**Perc.**
[0.-0.35)	17	17	100
[0.35-0.4)	55	54	98.2
[0.4-0.45)	79	78	98.7
[0.45-0.5)	107	106	99.1
[0.5-0.55)	137	136	99.3
[0.55-0.6)	198	192	97.0
[0.6-0.65)	120	116	96.7
[0.65-1.0)	32	30	93.8

In Figure 3, the cumulative probability distribution function of both series -recorded and simulated- are shown (data from Madrid).

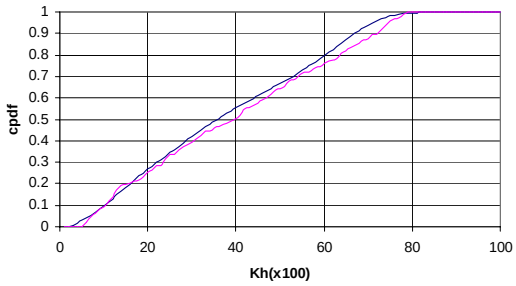

Figure 3. CPDFs for the recorded and simulated data. Interval 0.35--0.40. Madrid, 1979, January.

Moreover, we have calculated the hourly series of global irradiation from the hourly clearness index series. Using statistical tests, the hypothesis that both series have the same mean and variance is not rejected (significance level=0.05). The frequency histograms of the recorded and simulated series have been also analyzed. The frequency histograms have been obtained for each month of the year, using all the recorded and simulated series for that month over every year. The null hypothesis that the underlying model for both series is the same has never been rejected (significance level=0.05).

6. CONCLUSIONS

In this paper, a new model to predict climatic parameters is proposed. This model is based on the use of probabilistic finite automata and has been developed within the machine learning field. We have verified that this model allows us to keep all the relevant information contained in the univariate time series in an easy way. Moreover, with this mathematical model, the different relationships observed between different subsequences can be left out; each subsequence only uses the memory length that it requires.

Using this model, the series of a climatic parameter -global hourly radiation data- have been analyzed. First, a dynamic discretization method was used. Then, a set of PFAs -one for each interval of the analyzed parameter- were built. Using these PFAs, a method to predict new values of the parameter has been proposed.

Finally, the model has been checked using several tests. To check the cumulative probability distribution functions, an adaptable test, based on the two-sample Kolmogorov-Smirnov test, was used.

The obtained results prove that Probabilistic Finite Automata can be used to model climatic parameters and to predict new values.

6. REFERENCES

Box, G.E.P., Jenkins, G.M. Time series analysis forecasting and control. USA: Prentice-Hall, 1976.

Forbus, K.D. Qualitative process theory, *Artificial Intelligence*. 24: 85-168, 1984.

Kleer, J., Brown, J.S. A qualitative physics based on confluences. *Artificial Intelligence*, 24, 7-83, 1984.

Krog, A, Mian, S.I., Haussler, D. A hidden Markov model that finds genes in E.coli DNA. Technical report UCSC-CRL-93-16, University of California at Santa-Cruz, 1993.

Kuipers, B. Commonsense reasoning about causality deriving behavior from structure, *Artificial Intelligence*. 24, 169-203, 1984.

Mora-López, L., Fortes, I., Morales-Bueno, Triguero, F. Dynamic discretization of continuous values from time series. *Lecture Notes in Artificial Intelligence*, Vol. 1810, 280-291, 2000.

Nadas, A. Estimation of probabilities in the language model of the IBM speech recognition system. *IEEE Trans. on ASSP*, 32(4), 859-861, 1984.

Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the Seventh Annual Workshop on Computational Learning Theory, 1994.

Rissanen, J. A universal data compression system. *IEEE Trans. Inform. Theory*, 29(5), 656-664, 1983.

Rohatgi, V.K. "An Introduction to Probability Theory and Mathematical Statistics". John Wiley & Sons, USA, 1976.

Ron, D., Singer, Y., Tishby, N. Learning probabilistic automata with variable memory length. Proceedings of the Seventh Annual Workshop on Computational Learning Theory, 1994.

Ron, D., Singer, Y., Tishby, N. On the learnability and usage of acyclic probabilistic finite Automata. *Journal of Computer and System Sciences, 56*, 133-152, 1998.