Identification of Markov sequences based on a modified “forward-backward” algorithm

S.V. Shalagin[1], A.R. Nurutdinova[2]

[1]Kazan Federal University, Kremlevskaya street 18, Kazan, Russia, 420008
[2]National Research Technical University named after A.N. Tupolev, Karl Marks street 10, Kazan, Russia, 420111

Abstract. In 1989, L.R. Rabiner proposed a solution to the problem of identifying hidden Markov models by the induction method using a modified “forward-backward” algorithm. This article proposes the use of a modification of this method to solve the problem of identifying Markov sequences for belonging with a given probability to a specific class of the automate Markov model (AMM). A model defined on the basis of stochastic matrices of the class ergodic, which determine the class of the AMM.

1. Introduction

Markov models are used to solve various problems [2 – 5, 15, 24]. For example, displays of complex systems and processes in different areas [2-5]. In [6-17] one can find the results of studies of the analysis of the entropy and asymptotic properties of discrete Markov processes. A separate area of research in the theory of Markov processes consists of analysis tasks (recognition, control, and diagnostics) certain information about the structure or nature of the automaton for implementations that represent the automaton with a given accuracy [7, 8]. In [11, 12] methods for the classification and identification of Markov chains (MC), which arise in solving the recognition problems of automate Markov models (AMM), are proposed. In particular, they are based on the calculation of functions from sequences of finite length, taken from the AMM output, with a certain relative to the stochastic matrices (ESM) that defines it. This reduces the accuracy of the analysis based on the indicated functionals with restrictions on the length of the observed sequence, especially for the MC length of the order 10^2-10^3.

References [18, 19] consider issues related to AMM on the basis of the Markov chains generated by them. To solve the above problem, identification models were proposed that are based on the calculation of functionals directly from the implementation of the digital computer, taking into account the structure of the ESM that determines the AMM. According to the results, the methods obtained made it possible to increase the information content of the solution of the AMM recognition problem, to identify it with a higher confidence probability for a smaller number of generated elements of the CM.

In [20], an approach is proposed to the identification of automaton Markov models defined on the basis of ergodic stochastic matrices based on the implementations of Markov chains generated by them, which is a modification of the “forward-backward” algorithm proposed for solving the speech recognition problem for a hidden Markov model in [1]. The identification model is considered for two classes of automate Markov models defined on the basis of regular and cyclic ESM.

This paper shows the possibility of developing a modified "forward-backward" algorithm for Markov sequences, for which at a given time instant there is a subset of states observed with equal

probability. This algorithm can be applied, in particular, to describe and study objects from the field of quantum information processing [25].

2. The modification of the "forward-backward" algorithm.

We will call automate Markov models (AMM) as autonomous probabilistic automaton without outputs

$$(S, \phi(s'/ s)), \quad (1)$$

where $S = \{s_i\}$, $i = \overline{0, n-1}$– is a set of MC states, $s, s' \in S$, $\phi(s', s)$− is a transition function given by a stochastic matrix P_s, $P_s = (p_{ij})$ of $n \times n$, size , $i, j = \overline{0, n-1}$ [10]. If you set different functions $\phi(s', s)$ for AMM, you can get a finite set c of different subclasses of AMM, defined by the ESM P: $\{Q_k\}$, , $k = \overline{1, c}$, and to solve the problem of recognition of AMM from the implementation of experiments of MC generated based on given AMM subclasses. Subclasses of AMM are given depending on the structure of ESM P. The subclass of the ergodic stochastic matrix is determined by the location of the positive elements in its defined positions. In particular, in [11, 12] such subclasses of ESM as triangular upper, triangular lower, block right and block left are distinguished.

We introduce the following definitions.

Definition 1. $\hat{S}(N) = u_1, u_2, \ldots, u_N$ – the set of admissible implementations of experiments of a Markov chain given by an AMM of the form (1) at times t, where u_t – a subset of the states of the MC from the set S, admissible at the moment of time t, $t = \overline{1, N}$, with equal probability q^{-1}, $u_t \subset S$, $|u_t| = q \in [1, n]$.

Definition 2. Special case $\hat{S}(N) = u_1, u_2, \ldots, u_N$ – Markov chain, in which all states are completely observable: $|u_t| = 1$, $t = \overline{1, N}$.

Definition 3 For element $\hat{S}(N)$ is allowed to have a full set of implementations S at time t with equal probability n^{-1}, $u_t = S|u_t| = n$, which corresponds to a completely unobservable state of the MC.

It is necessary to determine the value $P(\hat{S}(N)|\text{AMM}(P))$ is the probability that the set $\hat{S}(N)$ generated on the basis of AMM P, where the ESM P belongs to a given subclass Q_k.

For identification according to the "forward-backward" algorithm of the fact that set $\hat{S}(N) = u_1, u_2, \ldots, u_N$ is generated on the basis of AMM ($P \in Q_k$) for given k, the following variable arrays are introduced [1]: $\alpha_t(i) = P(u_1, u_2, \ldots, u_t, u_t = s_i|\text{AMM}(P))$, $t = \overline{1, N}$, $i = \overline{1, m}$.

The identification algorithm for an automate Markov model of the form (1) based on the particular case of a set $\hat{S}(N) = u_1, u_2, \ldots, u_N$, defined according to definition 2, consists of the following steps [4].

Step 1. Initialization of variables: $\alpha_1(i) = \pi_0(i) \cdot z_i(1)$, $z_i(1) = \begin{cases} 1: & u_1 = s_i \\ 0: & \text{otherwise} \end{cases}$, $i = \overline{1, n}$.

Step 2. Induction: $\alpha_{t+1}(j) = \left(\sum_{i=1}^{n} \alpha_t(i) \cdot p_{ij}\right) \cdot z_j(t+1)$, $z_i(t+1) = \begin{cases} 1: & u_{t+1} = s_i \\ 0: & \text{otherwise} \end{cases}$, $t = \overline{1, N-1}$, $j = \overline{1, n}$.

Step 3. Find the value $P(\hat{S}(N)|\text{AMM}(P)) = \alpha_N(s(N))$.

Consider the case when the full set of implementations is valid for at least one element $\hat{S}(N) = u_1, u_2, \ldots, u_N$ by definition 3. According to [25], at the stage of calculating the values $\alpha_{t+1}(i)$, $t = \overline{1, N-1}$, $i = \overline{1, n}$, the expression takes place:

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{n} \alpha_t(i) \cdot p_{ij}\right) \cdot z'_j(t+1), z'_i(t+1) = \begin{cases} 1: & |u_{t+1}| = n \\ z_i(t+1): & \text{otherwise} \end{cases}.$$

If in the observed sequence $s(N)$ there are k elements for which $|u_t| = n$, then the required probability is

$$P\left(\hat{S}_k(N)|\text{AMM}(P)\right) = \sum_{i=1}^{n} \alpha_N(i). \quad (2)$$

The computational complexity of the proposed method for solving the problem of identifying finite simple homogeneous Markov chains has order $O(N \cdot n)$ if all its elements are observable. The presence of hidden elements in an amount comparable to a long sequence N increases the order of the computational complexity of the algorithm to $O(N \cdot n^2)$ [21].

In references [22-23] estimates of the complexity of algorithms for identifying finite simple homogeneous Markov chains were calculated. In particular, it was shown that using functional based on l-grams, $l = 2,3$, the order of computational complexity of the identification algorithm is equal,

$O(N \cdot n^2)$ and $O(N \cdot n^3)$ accordingly, it is therefore advisable to use these methods for small values of n. For large values of n, it is more efficient to use an algorithm based on frequency features, the computational complexity of which is of the order $O(N \cdot n)$ or the proposed "forward-backward" algorithm. The proposed model allows us to quantify the probability of identifying the sequence of a Markov chain in terms of the possibility of generating a given AMM.

In general, to identify a set $\hat{S}(N) = u_1, u_2, \dots, u_N$, the development of the "forward-backward" algorithm is as follows.

Step 1. Initialization of variables: $\alpha_1(i) = \pi_0(i) \cdot z_i(1)$, $z_i(1) = \begin{cases} q_1^{-1}: & s_i \in u_1 \\ 0: & otherwise \end{cases}$, $q_1 = |u_1|$, $i = \overline{1, n}$.

Step 2. Induction: $\alpha_{t+1}(j) = \left(\sum_{i=1}^{n} \alpha_t(i) \cdot p_{ij}\right) \cdot z'_j(t+1)$, $z'_j(t+1) = \begin{cases} n^{-1}: & |u_{t+1}| = n \\ z_j(t+1): & otherwise' \end{cases}$

$z_j(t+1) = \begin{cases} q_{t+1}^{-1}: & s_j \in u_{t+1} \\ 0: & otherwise \end{cases}$, $q_{t+1} = |u_{t+1}|$, $j = \overline{1, n}$, $t = \overline{1, N-1}$.

Step 3. Find the value $P(\hat{S}(N)|\text{AMM}(P)) = \alpha_N(s(N))$ according to (2).

We define the complexity of the proposed algorithm. At stage 1, provided that $|u_1| > 1$, the $|u_1|$ operations of multiplication by a constant $(\otimes_{cn})\, q_1^{-1}$. Otherwise, only the assignment operation of the form $\alpha_1(i) = \pi_0(i)$. In step 2, for $t = \overline{1, N-1}$, if $q_{t+1} = 1$, then perform n^2 of multiplication operations $(\otimes)$, $n(n-1)$ addition operations $(\oplus)$; если $q_{t+1} > 1$, then additionally perform $n \cdot q_{t+1}$ operations $\otimes_{cn} q_{t+1}^{-1}$. In step 3 $(n-1) \oplus$ are performed. Let be $d_t = \begin{cases} q_t: & q_t > 1 \\ 0: & \text{иначе} \end{cases}$, $t = \overline{1, N}$.

Statement. The computational complexity of the proposed algorithm by the number of multiplication, addition and multiplication operations by a constant is : $(N-1)n^2$, $(n-1)(n(N-1)+1)$ ang $d_1 + n \cdot \sum_{t=1}^{N-1} d_{t+1}$, respectively.

According to the statement, the computational complexity of the proposed algorithm has an order $O(N \cdot n^2)$ in quantity $\otimes$ and quantity $\oplus$. Depending on the type of the identified set $\hat{S}(N)$ the complexity of the algorithm differs only in the number of operations of multiplication by a constant. The number of these operations is determined by the power of the subset the complexity of the algorithm differs only in the number of operations of multiplication by a constant. The number of these operations is determined by the power of the subset. $q_{t+1} = |u_{t+1}| \leq n$.

3. Conclusion

Thus, the proposed modified "forward-backward" algorithm is quite effective in assessing the computational complexity in comparison with methods using functions based on l-grams. In addition, the method allows to solve the recognition problem for sequences with hidden and partially identified elements. The complexity of the algorithm in the number of operations of multiplication and addition does not depend on the degree of certainty (or uncertainty) of the elements of the identifiable set, but varies only in the number of operations of multiplication by a constant. It is important that the number of operations increases linearly with increasing size and quadratically with increasing size of matrices, on the basis of which recognizable values can be identified.

4. References

[1] Rabiner, L.R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition / L.R. Rabiner // Proc. IEEE. – 1989. – Vol. 77(2). – P. 257-286.

[2] Levin, B.R. Probabilistic models and methods in communication and management systems / B.R. Levin, V. Shvarts – M.: Radio and Communications, 1985. – 312 p.

[3] Friedman, W.F. Military cryptanalis / W.F. Friedman – Aegean Park Press, Laguna Hills CA. – 1985. – Vol. Z(1). – P. – 342 p.

[4] Teptin, G.M. Markov models of means of protection of the automated systems of special purpose / G.M. Teptin, K.V. Ivanov // Uchen. zap. Kaz. un-ta. Ser. Phys.-mod. sciences. – 2008. – Vol. 150(4). – P. 41-53.

[5] Raskin, L.G. Analysis of stochastic systems and elements of optimal control theory / L.G. Raskin – M.: Soviet Radio. – 1976. – 344 p.
[6] Pospelov, D.A. Probabilistic automata / D.A. Pospelov – M.: Energy, 1970. – 88 p.
[7] Bukharaev, R.G. To the problem of minimization of the input automaton that generates a given homogeneous Markov chain / R.G. Bukharaev // Kazan. zap. Kaz. Un-ty. Ser. phys.-mat. sciences. – 1967. – Vol. 129(4). – P. 3-11.
[8] Bukharaev, R.G. The representability of events in probabilistic automata / R.G. Bukharaev // Kazan. zap. Kaz. Un-ty. Ser. Phys.-mat. –1967. – Vol. 127(3). – P. 7-20.
[9] Alpin, Y.A. On the normal form of a stochastic matrix / Y.A. Alpin // Kazan. zap. Kaz. Un-ty. Ser. phys.-mat. – 2012. – Vol. 154(2). – P. 60-72.
[10] Bukharaev, R.G. Probabilistic automata / R. G. Bukharaev – Kazan: Publishing house of KSU. − 1970. − 188 p.
[11] Zakharov, V.M. Analysis of stochastic matrices by multivariate classification / V. M. Zakharov, N.N. Nurmeev, F.I. Salimov // Discrete mathematics and its applications: proceedings of the 7th Intern. seminar. – 2001. – Vol. 3. – P. 156-159.
[12] Zakharov, V.M. Classification of stochastic ergodic matrices of cluster and discriminant analysis methods / V.M. Zakharov, N.N. Nurmeev, F.I. Salimov // Studies in Informatics. – 2000. – Vol. 2. – P. 91-106.
[13] Lorentz , A.A. Reliability and speed of probabilistic automata / A.A. Lorentz – Riga: Zinatne, 1976. – 112 p.
[14] Romanovsky, V.I. Discrete Markov chains / V.I. Romanovsky – Moscow: Gostekhizdat, 1949. – 436 p.
[15] Fedotov, N.G. Methods of stochastic geometry in image recognition / N.G. Fedotov – M.: Radio and communication, 1990. – 144 p.
[16] Kemeny, J. Finite Markov chains / J. Kemeny, J. Snell. – Moscow: Science, 1970. – 272 p.
[17] Lee, I. Estimating the parameters of Markov models for aggregated time series / I. Lee, A. Zellner – M.: Statistics, 1977. – 221 p.
[18] Nurutdinova, A.R. Methodology identification of automaton Markov models based on the resulting sequence / A.R. Nurutdinova, S.V. Shalagin // Herald of KSTU named after. A. N. Tupolev. – 2010. – Vol. 1. – P. 94-99.
[19] Nurutdinova, A.R. In multi-parametric classification of automaton Markov models based on generated sequences of their States / A.R. Nurutdinova, S.V. Shalagin // Applied discrete mathematics. – 2010. – Vol. 4. – P. 41-54.
[20] Shalagin, S.V. Identification of Markovian Automata Sub-classes // International Journal of Pharmacy and Technology / S.V. Shalagin, A.R. Nurutdinova. – 2016. – Vol. 8(3). – P. 15327-15337.
[21] Shalagin, S.V. Identification Algorithms of Simple Homogeneous Markov Chains of Cyclic Class and Their Complexity Analysis // S.V. Shalagin, A.R. Nurutdinova // International Journal of Pharmacy and Technology. – 2016. – Vol. 8(3). – P. 14926-14935.
[22] Shalagin, S.V. Identification of the sequence of measurements of economic parameters on the basis of the hidden Markov model / A.R. Nurutdinova, S.V. Shalagin // Problems of analysis and modeling of regional socio-economic processes: proceedings of reports VII Inter. full-time science.- prakt. Conf. – Kazan: Publishing house of KSU, 2017. – P. 159-162.
[23] Nurutdinova, A.R. Identification of automaton Markov models using a modified "forward-reverse" algorithm / A.R. Nurutdinova // Control systems and information technologie. – 2018. – Vol. 2(72). – P. 36-41.
[24] Raikhlin, V.A. Reliable Recognition of Masked Binary Matrices / V.A. Raikhlin, I.S. Vershinin, R.F. Gibadullin, S.V. Pystogov // Connection to Information Security in Map Systems, Lobachevskii Journal of Mathematics. – 2013. – Vol. 34(4). – P. 319-325.
[25] Ablayev, F.M. Multi-qubit controlled NOT gates for artificial intelligence natural languages processing / F.M. Ablayev, S.N. Andrianov, N.S. Andrianova, A.A. Kalachev, A.V. Vasiliev // Proceedings of SPIE-The International Society for Optical Engineering (International Conference on Micro- and Nano-Electronics). – 2019. – Vol. 11022. – P. 110222B.