Information Sciences 326 (2016) 188–201

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Recursive estimation of high-order Markov chains: Approximation by finite mixtures

Miroslav Kárný*

Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 182 08 Prague, Czech Republic

ARTICLE INFO

Article history:
Received 23 February 2014
Revised 29 June 2015
Accepted 17 July 2015
Available online 23 July 2015

Keywords:
Markov chain
Approximate parameter estimation
Bayesian recursive estimation
Adaptive systems
Kullback–Leibler divergence
Forgetting

ABSTRACT

A high-order Markov chain is a universal model of stochastic relations between discrete-valued variables. The exact estimation of its transition probabilities suffers from the curse of dimensionality. It requires an excessive amount of informative observations as well as an extreme memory for storing the corresponding sufficient statistic. The paper bypasses this problem by considering a rich subset of Markov-chain models, namely, mixtures of low dimensional Markov chains, possibly with external variables. It uses Bayesian approximate estimation suitable for a subsequent decision making under uncertainty. The proposed recursive (sequential, one-pass) estimator updates a product of Dirichlet probability densities (pds) used as an approximate posterior pd, projects the result back to this class of pds and applies an improved data-dependent stabilised forgetting, which counteracts the dangerous accumulation of approximation errors.

1. Introduction

Adaptive systems select their actions while simultaneously learn dynamics of the environment they interact with [5]. The joint acting and learning is their key feature, which singles them from other controllers, estimators, predictors, classifiers, etc. It is the main source of their strength, applicability [40] and universality [27]. It allows them to work with simple, often input–output, models describing their interactions with their environments locally [23]. Recursively estimated black-box models, relating the observed past to future observations, serve to adaptive systems exploited for prediction [4], decision support [42], feedback control [43], etc.[1] All of them de facto perform dynamic decision making under uncertainty, which has Bayesian paradigm [10,49] as the theoretically most elaborated base. This makes us to focus on it.

Often, the modelled observations as well as explanatory variables are discrete or discretised. Then, a high-order Markov chain provides their universal model. It relates the predicted observation to a finite-dimensional regression vector containing the past observations and explanatory variables. Its recursive Bayesian estimation, which provides a lossless compression of the available knowledge, is formally simple. Basically, it counts joint occurrences of the predicted variable and the corresponding regression vector. The applicability is, however, strongly limited by the curse of dimensionality [8] as the size of the occurrence array blows up with the number of possible occurrence instances. Then, the observations insufficiently populate it.

* Tel.: +420266052274.
E-mail address: school@utia.cas.cz

[1] The vast literature on the topic forces us to provide just subjectively selected samples.

http://dx.doi.org/10.1016/j.ins.2015.07.038
0020-0255/

This curse is mostly counteracted by exploiting conditional independence of modelled variables, cf. Bayesian network [21], compositional models [22], etc. This quite powerful way allows coping even with computationally hard designs of decision strategies [12,19], which exploit specific factored representations. The data-based factorisation into conditionally independent groups itself faces, even more pronounced, curse of dimensionality. This limits universality of these techniques and especially their use in exploratory data analysis and black-box-models-based adaptive systems.

The current paper deals with a rich class of estimation problems in which a detailed independence structure is not supplied by experts. The considered regression problems are delimited by a few possible observation values but a long regression vector is allowed. For them, a finite mixture of Markov chains relating the predicted observation to individual entries of the regression vector is employed. Demonstrations of the modelling strength of such a mixture are in [9,45,48]. The employed solutions, however, fully rely on batch processing, which prevents their permanent use to adaptive systems, data stream processing and modelling of non-stationary phenomena [47].

The recursive learning of a finite mixture of Markov chains is inevitably approximate and consequently endangered by accumulation of approximation errors. In parameter estimation, the "natural" error-damping effect of state-space (partially observable) models [13] does not exist and the accumulation-errors-free algorithms [36,37] operate on too narrow class of non-sufficient statistics. A recent inspection of this problem shown that stabilised forgetting [25] provides a general counter-measure against the discussed accumulation.

The stabilised forgetting and handling of Dirichlet pds summarised [26] provide essential ingredients for the inspected *approximate Bayesian recursive estimation of finite Markov-chain mixtures*.

The estimator design is addressed within the unified Bayesian paradigm [11,29], which respects its use in a subsequent Bayesian decision making. Practically, it makes us to "neglect" important techniques, which are of a heuristic nature or less rigorously connected with the decision making. They include classical quasi-Bayes estimator [52] and its dynamic version [26] (heuristic), variational Bayes [51] (improper order of arguments in minimised Kullback–Leibler divergence), expectation propagation [41] or various point estimators (information on precision is missing), often based on expectation-maximisation algorithm [17] possibly combined with variational approach [50]. In summary, the paper:

- fills the niche in an extreme range of variants of estimating of models exploiting Markov-chains and complements toolkit of estimators dealing with mixed and/or hidden variables;
- adopts parsimonious probabilistic model of finite-mixture type [9,45] relating discrete-valued predicted variable to its multiple delayed values and to external discrete-valued variables;
- designs novel recursive Bayesian estimator of this black-box input–output model, which fits to its intended use for adaptive decision making operating solely on discrete-valued observations;
- designs approximate, recursive estimator via a recent systematic, theoretically justified, way [25]: this distinguishes the estimator from its heuristically justified predecessors like quasi-Bayes estimator [26,52];
- serves as an example of the general theory of approximate recursive estimation [25] and improves its only heuristic step encountered.

Concerning the layout, Section 2 formalises the problem. Core Section 3 solves it. Section 4 describes how to counteract accumulation of errors caused by the approximate estimation and improves the heuristic step adopted in [25]. Section 5 illustrates the theory and its notions on a simple case of exploratory analysis. Other examples there provide comparison with a few published examples addressing similar objectives as the current paper. Section 6 adds concluding remarks.

2. Formalisation of the addressed problem

Scalar[2] discrete-valued *observations* $\Delta_t \in \boldsymbol{\Delta} \equiv \{1,\ldots,|\boldsymbol{\Delta}|\}$, $|\boldsymbol{\Delta}| \ll \infty$ are made at discrete time moments $t \in \boldsymbol{t} \equiv \{1,\ldots,T\}$, $T \leq \infty$, on the modelled stochastic environment. The observation Δ_t is assumed to depend on an $\ell_{\boldsymbol{\psi}}$-dimensional *regression vector* $\psi_t \in \boldsymbol{\psi}$ with discrete-valued entries[3] $\psi_{t;i} \in \boldsymbol{\psi}_i \equiv \{1,\ldots,|\boldsymbol{\psi}_i|\}$, $|\boldsymbol{\psi}_i| \ll \infty$, $i \in \boldsymbol{i} \equiv \{1,\ldots,\ell_{\boldsymbol{\psi}}\}$. The unknown dependence is assumed to be time-invariant. Thus, the observations are described by the *high-order Markov chain*, parameterised by time-invariant finite-dimensional array of transition probabilities Ω,

$$\mathrm{p}(\Delta_t|\Omega,\text{observed past}) = \mathrm{p}(\Delta_t|\Omega,\psi_t) \equiv \prod_{\psi\in\boldsymbol{\psi}}\prod_{\Delta\in\boldsymbol{\Delta}} \Omega_{\Delta|\psi}^{\delta(\Delta\Delta_t)\delta(\psi\psi_t)}. \tag{1}$$

There, p denotes *probability density (pd)* of the variable in its argument conditioned on the argument after the conditioning symbol |. Ω contains time-invariant unknown probabilities belonging to an appropriate probabilistic simplex. The subscript $_{\Delta|\psi}$ of its entries stresses that $\Omega_{\Delta|\psi}$ is the transition probability with properties

$$\Omega_{\Delta|\psi} \geq 0,\ \forall\Delta\in\boldsymbol{\Delta},\ \forall\psi\in\boldsymbol{\psi} \quad \text{and} \quad \sum_{\Delta\in\boldsymbol{\Delta}} \Omega_{\Delta|\psi} = 1\ \ \forall\psi\in\boldsymbol{\psi}.$$

[2] The consideration of a scalar-valued observation represents no restriction as the vector case can always be boiled down to it by employing entry-wise modelling [26]. $\boldsymbol{x}$ denotes a set of x values, $|\boldsymbol{x}|$ its cardinality and $\equiv$ means defining equality. Section 5 illustrates the notions being introduced.

[3] The symbol $\ell_{\boldsymbol{x}}$ is reserved for the length of a vector $x \in \boldsymbol{x}$. The time index $t \in \boldsymbol{t}$ is always the first one and semicolon separates it from other indices. The time index indicates that the values of observed and predicted variables or statistic values are meant.