Robust Estimation of Context Trees (Simulations)\footnote{This work is supported by CAPES}

Márcio Luis Lanfredi Viola
Jesus Enrique Garcia
Verónica Andrea González-López

Words key: Context Tree, Context Algorithm, Variable Memory Markov Chain, CTM Algorithm, Relative Entropy, Robustness.

1 Robust Estimator of Context Trees

Let $A$ a finite alphabet. The string $s = a_m a_{m+1} \ldots a_n$, $a_i \in A$, $m \leq i \leq n$ is denoted by $a^n_m$ and its length is represented by $l(s)$.

A set $\tau$ of strings is represented by a postfix tree, that is, if no $s_1 \in \tau$ is a postfix of any other $s_2 \in \tau$. The depth of the tree $\tau$ is $d(\tau) = \max\{l(s), s \in \tau\}$.

Each string $s = a^k_1 \in \tau$ is visualized as a path from a leaf to the root (drawn with the root at the top), consisting of $k$ edges labeled by symbols $a_1 \ldots a_k$. The strings $s \in \tau$ are identified also with the leaves of the tree $\tau$, leaf $s$ is the leaf connected with the root by the path visualizing $s$ as above. Similarly, the nodes of the tree $\tau$ are identified with the finite postfixes of all (finite or infinite) $s \in \tau$, the root being identified with the empty string $\emptyset$. The children of a node $s$ are those strings $as$, $a \in A$, that are themselves nodes, that is, postfixes of some $s' \in \tau$ [3, 6].

Let us be given a stationary ergodic stochastic process $\{X_i, -\infty < i < \infty\}$ with finite alphabet $A$. Write $Q(a^n_m) = P(X^m_n = a^n_m)$ and, if $s \in A^k$ has $Q(s) > 0$, write $Q(a|s) = P(X_o = a|X^{-1}_{-k} = s)$. A process as this will be referred to as process $Q$.

A string $s \in A^k$ is a context for a process $Q$ if $Q(s) > 0$ and $P(X_o = a|X^{-1}_{-k} = x^{-1}_{-\infty}) = Q(a|s)$, for all $a \in A$ whenever $s$ is a postfix of the sequence $x^{-1}_{-\infty}$, and no proper postfix of $s$ has this property. Clearly, the set of all contexts is a tree and it will be called the context tree of the process $Q$ [1, 3, 6].

Considering $m$ samples, if some kind of contamination affect any of these samples, the structure of the estimated tree could change. If we have only one sample, we can split it into $m$ samples.

We consider $m$ independent samples (strings). Each sample come from one of two possible Variable Memory Markov Chain [1] with context trees $\tau_Q$ or $\tau_{Q'}$, respectively. Each sample is generated from tree $\tau_Q$ with probability $\gamma$ or tree $\tau_{Q'}$ with probability $(1-\gamma)$, $1/2 < \gamma < 1$, that is, we consider the mixture model $\gamma \tau_Q + (1-\gamma) \tau_{Q'}$, $1/2 < \gamma < 1$.\"
Considering $m$ independent samples $x_1^n, ..., x_m^n$, we can estimate a tree $\tau_i, i = 1, ..., m$, through the algorithm proposed by Csiszár & Talata [3] for each sample. Moreover, for each two trees $i$ and $j$, $i = 1, ..., m, j = 1, ..., m$, we calculate $D(\tau_i||\tau_j)$ where $D(\tau_i||\tau_j)$ is entropy rate [2] of the chain corresponding to the tree $\tau_i$ with respect to the chain corresponding to the tree $\tau_j$. The calculations $D(\tau_i||\tau_j)$ are based in the following theorem.

**Theorem 1.** Let $A$ an finite alphabet. Consider two Variable Memory Markov Chain given by trees $\tau_P$ and $\tau_Q$, respectively, with law $P$ and $Q$. Then, the entropy rate of the chain corresponding to the tree $\tau_P$ with respect to the chain corresponding to the tree $\tau_Q$ is given by

$$D(\tau_P||\tau_Q) = \sum_{s \in S} P(s)D(P(.|s)||Q(.|s))$$

where $S = \{s \in \tau_P \cup \tau_Q : \forall s' \in \tau_P \cup \tau_Q, s = s' \text{ se } s \text{ é sufixo de } s'\}$.

We can choose a context tree $\hat{\tau}_j$ by following procedure:

**Definição 1.** Consider an sample $x_i^n, i = 1, ..., m$, and their respective estimate tree $\hat{\tau}_1, ..., \hat{\tau}_m$. The proposed estimator is

$$\text{arg} \min_{i = 1, ..., m} \left\{ \sum_{i = 1}^{m} D(\hat{\tau}_i||\hat{\tau}_1), ..., \sum_{i = 1}^{m} D(\hat{\tau}_i||\hat{\tau}_m) \right\} \quad (1.1)$$

where $D(\|\|.)$ is the entropy rate between two context trees given by Theorem 1.

We calculate the finite sample addition breakdown point and the finite sample replacement breakdown-point [5, 7] of the proposed estimator above. The finite sample addition breakdown point and the finite sample replacement breakdownpoint were proposed by Donoho e Huber [4].

Consider the following conditions:

- $\exists c_1$ such that $\text{Prob}(D(\hat{\tau}_a||\hat{\tau}_c) < c_1) \geq 1 - \epsilon$;
- $\exists c_2$ such that $\text{Prob}(D(\hat{\tau}_b||\hat{\tau}_d) < c_2) \geq 1 - \epsilon$;
- $\exists M$ such that $\text{Prob}(D(\hat{\tau}_a||\hat{\tau}_b) < M) \geq 1 - \epsilon$;

where $\hat{\tau}_a, \hat{\tau}_c$ were the estimated trees of the samples from Variable Memory Markov Chain with context tree $\tau_Q$, and $\hat{\tau}_b, \hat{\tau}_d$ where the estimated trees of the samples from Variable Memory Markov Chain with context tree $\tau_Q$. Let $c = \max\{c_1, c_2\}$.

**Theorem 2.**

1. The finite sample addition breakdown point of the estimator (1.1) is $\frac{1}{2}$.

2. The finite sample replacement breakdownpoint of the estimator (1.1) is $\frac{1}{2}$.

We show the proof of the Theorems 1 and 2 and we show four scenarios simulation.
2 Simulations

2.1 Simulation 1:

Consider the context trees:

- $\tau_1$: P(0|0)=0.8, P(0|1)=0.7, P(0)=0.78;
- $\tau_2$: P(0|0)=0.75, P(0|1)=0.7, P(0)=0.73;
- $\tau_3$: P(0|0)=0.8, P(0|1)=0.75, P(0)=0.785;
- $\tau_4$: P(0|0)=0.75, P(0|1)=0.75, P(0)=0.75;
- $\tau_5$: P(0|0)=0.2, P(0|1)=0.298, P(0|01)=0.3, P(0|11)=0.3, P(0)=0.261, P(01)=0.2192, P(11)=0.5145.

The tree $\tau_5$ has contamination in its structure to choose one of five trees $\tau_i, i = 1, ..., 5$.

2.2 Simulation 2:

We simulate four samples of length 3000 from process with alphabet $\mathcal{A} = \{0, 1, 2\}$ and context tree 002, 22, 1, 0, 102, 202 e 12. Moreover, we simulate one sample of length 3000 from process with context tree 112, 22, 212, 1, 0, 012 e 02 (this context tree has contamination in its structure).

We use the algorithm proposed by Csiszár & Talata [3] to estimate the context trees from each sample and the proposed robust procedure 1.1 is applied to choose one of the estimated trees $\hat{\tau}_i$.

2.3 Simulation 3:

Consider the context trees:

- $\tau_1$: P(0|0)=0.8, P(0|1)=0.7, P(0)=0.77;
- $\tau_2$: P(0|0)=0.75, P(0|1)=0.7, P(0)=0.73;
- $\tau_3$: P(0|0)=0.8, P(0|1)=0.75, P(0)=0.785;
- $\tau_4$: P(0|0)=0.75, P(0|1)=0.75, P(0)=0.75;
- $\tau_5$: P(0|0)=0.2, P(0|1)=0.3, P(0)=0.27.

The tree $\tau_5$ has contamination in its structure.

The proposed robust procedure 1.1 is applied to choose one of five trees $\tau_i, i = 1, ..., 5$. 
2.4 Simulation 4:

This simulation is divided into four stages:

The first step, we simulate five samples of length 15000. The first four simulations from process with alphabet $\mathcal{A} = \{0, 1\}$ and context tree 00, 10, 01 e 11 with transition probabilities $P(0|00)=0.8$, $P(0|10)=0.6$, $P(0|01)=0.4$ and $P(0|11)=0.7$. The fifth simulated sample from process with context tree 00, 10, 01 e 11 with transition probabilities $P(0|00)=0.2$, $P(0|10)=0.5$, $P(0|01)=0.4$ and $P(0|11)=0.3$ (this context tree has contamination in its transition probabilities).

In step 2, we replace the fourth simulated sample for an sample from process with transition probabilities $P(0|00)=0.2$, $P(0|10)=0.5$, $P(0|01)=0.4$ e $P(0|11)=0.3$.

In step 3, we replace the third simulated sample for an sample from process with transition probabilities $P(0|00)=0.2$, $P(0|10)=0.5$, $P(0|01)=0.4$ e $P(0|11)=0.3$.

In step 4, we replace the second simulated sample for an sample from process with transition probabilities $P(0|00)=0.2$, $P(0|10)=0.5$, $P(0|01)=0.4$ e $P(0|11)=0.3$.

We use the algorithm proposed by Csiszár & Talata [3] to estimate the context trees and, in each step, the proposed robust procedure 1.1 is applied to choose one of the estimated trees $\hat{\tau}_i$. 
Referências


