Entropy-Rate Driven Inference of Stochastic Grammars

Unto K. Laine

Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland

Abstract

A new method for inferring specific stochastic grammars is presented. The process called Hybrid Model Learner (HML) applies entropy rate to guide the agglomeration process of type ab→c. Each rule derived from the input sequence is associated with a certain entropy-rate difference. A grammar automatically inferred from an example sequence can be used to detect and recognize similar structures in unknown sequences. Two important schools of thought, that of structuralism and the other of ‘stochasticism’ are discussed, including how these two have met and are influencing current statistical learning methods. It is argued that syntactic methods may provide universal tools to model and describe structures from the very elementary level of signals up to the highest one, that of language.

Index Terms: entropy, stochastic grammars, grammatical inference, pattern discovery, stop consonant classification

1. Introduction

The birth of the formal language theory in the 50’s inspired many branches of information related sciences [1]. Syntactic methods soon found their way also to the field of pattern recognition [2]. These were new phases in a long line concerning the evolution of structuralism (Ferdinand de Saussure, 1857-1913). By definition, patterns in reality are structured. They are constructed of certain parts called “primitives” or elements. “Constructed” means that the elements are organized functionally and/or structurally in an efficient and purposeful way. An entity constructed by randomly selected and connected elements seldom forms an important structure or pattern.

Structural aspects are essential in all human cognitive processes. Foremost, we hear and see, produce and remember, structures. Structures can be described in a compact form, which helps to recognize and remember them [3]. From this point of view it is not surprising that the most brilliant product of human cultural history – language – is highly structured from individual speech sounds up to concepts and syntax.

1.1. Inductive and grammatical inference

By induction we (or computational agents) try to formulate plausible general assertions that explain the given specific observations (facts) and are able to predict new observations (facts). Almost all learning mechanisms involve some kind of inductive inference. Two subclasses are typically mentioned: 1) learning from observations, and 2) learning from examples.

Induction and deduction are the two main methods to cope with material reality, see Figure 1. Inductive inference is always occurring in some context, called background knowledge. By induction we construct hypotheses on the information given by the observations of the real world. Hypotheses can be called models because they try to explain and summarize the observations.

The material reality closes the induction-deduction loop. Every artifact or construction tests the theory and models related to it. A bad theory may lead to a realization that doesn’t work. By observing the results of implementations we gain new information for continuous learning and model improvements. The learning loop is closed.

Based on this general view we could argue that natural languages provide tools to construct “natural models” (verbal models) of reality. Grammatical inference can be seen as a special case of inductive inference. It constructs grammatical models (rules) for the sentence structures of natural languages.

As already mentioned, syntactic methods have been applied for quite a long time in pattern recognition. However, grammars of natural languages may reflect some deeper, universal principles from which even our cognitive skills have emerged. Since we are able to learn those structures and we communicate through them, they must have a special role in relation to our cognitive system. On the other hand, the neocortex (where important parts of cognitive processing takes place) is functionally surprisingly universal and plastic; neural structures at different locations may learn the same task, if necessary. Based on these facts it may appear that syntactically (grammatically) describable structures play a much wider role in all cognitive processes than what has been earlier understood. They may even be involved in latent bottom-up pattern learning, discovering, modeling, memorizing and recognition mechanisms. In any case, the syntactic approach should be extended to study, model and simulate even these cognitive processes.

These general views and ideas (whether correct or incorrect) motivated this work, where new, universal and robust methods for grammatical inference were studied with focus on noisy sequential patterns. Specifically, we wanted to test new syntactic methods applied to raw data sequences obtained directly from speech signals by simple operations like permutation transformation [4].

This paper briefly introduces one outcome of the ongoing research: Hybrid Model Learner (HML) - a new method for sequential pattern discovery, learning and modeling. It applies universal information theoretical measures (especially entropy rate) to perform grammatical inference. It automatically (also in unsupervised manner) learns models by inferring specific
stochastic context-free grammars (SCFG) for patterns in discrete, noisy sequences and applies the found compact (grammatical) models to recognize similar structures in unknown sequences.

The main goal of this paper is to introduce the HML method, discuss its background, and to illuminate its main principles and properties. Its performance is demonstrated by an example from speech research: clustering of stop consonant bursts directly in the time domain.

2. Randomness and statistical thermodynamics vs. structuralism

‘Stochasticism’ integrates views related to change, noise, randomness and chaos; it has evolved quite independently of structuralism. These two contrasting views and aspects of reality (static structures vs. variability) seem to work together in many ‘lively’ emergent processes, including life itself [5]. One way to characterize these two entities is the observation that robust structures provide high predictability (e.g., in sequences), whereas randomness reduces it. Venn was one of the first to discuss the concept of randomness scientifically [6]. Around that time the concept of thermodynamics started to develop in physics taking ground from the dominating mechanistic (deterministic) Newtonian世界观. Progress in physical thermodynamics was needed in order to improve the extremely inefficient steam engines of the early 19th century.

Clausius coined the new concept of thermodynamic systems around 1850 and later he published an equation that can be interpreted as a precursor formulation of entropy. In 1865 he finally proposed the concept of entropy for his new ‘energy related measure’. In 1877 Ludvig Boltzmann formulated an alternative, statistical definition for entropy:

\[ S = k_B \log \Omega \]  

(1)

where \( k_B \) is a constant and \( \Omega \) the number of (non-distinguishable) microstates consistent with the given macrostate. This work formed a cornerstone for new branches in science: statistical thermodynamics and later statistical mechanics.

2.1. Shannon entropy, mutual information, etc.

Seventy years later Claude Shannon borrowed the idea of thermodynamic entropy and produced a communication theoretic interpretation of it [7]. This variant is called Shannon entropy or, information entropy, and is defined as:

\[ h(X) = -\sum_{k=1}^{K} p_k \log_2 p_k \]  

(2)

where \( h(X) \) denotes the information entropy of the sequence \( X \) and \( p_k \) the probability of symbol (alphabet, element) \( k \) in \( X \) constructed of \( K \) different symbols. We should remember that this measure is not true entropy. Firstly, its dimension is not energy but bit/symbol (the Boltzmann constant is missing). Secondly, the total information entropy \( H \) of the sequence (in bits) is \( h \times L \) where \( L \) is the length of the sequence. Therefore, \( h \) is called entropy rate.

Shannon entropy is often used to measure the randomness of a sequence. In order for (2) to be valid two premises must exist: the sequence must be infinite in length and occurrences of the symbols must be independent from each other (no constraints). It is easy to see that typically neither of these premises is true in a practical case. Therefore, more sophisticated methods have been developed to perform entropy (and randomness) estimation for finite sequences.

These questions are considered in algorithmic information theory together with themes like algorithmic complexity and algorithmic entropy. Briefly, (2) typically gives a too large value when the occurrences of the elements in the sequence have some constraints [8, 9]. In the MaxEnt principle [10] the missing probabilities of some distribution can be estimated with minimally biased values by applying the Shannon entropy. Mutual information is also a part of the Shannonian tradition. It was applied in grammatical inference by Wong et al. [11].

2.2. Two schools of thought meet

Shannon was one of the first to apply statistical methods and theories to the field of linguistics. He even made the first estimates of the entropy (rate) of English text. Zellig Harris, with a background in structuralism, also studied the application of statistical methods to linguistics. His student, Noam Chomsky, was not very happy with this new approach. Some authors have made critical comments related to the known split in the 60’s between these two scientists: “Chomsky’s influential remarks had the effect of killing off interest in probabilistic methods for syntax, just as for a long time McCarthy and Hayes (1969) discouraged exploration of probabilistic methods in Artificial Intelligence” [12]. However, currently these two long and wide schools of thought continue to cope with each other in a deep and rich manner more than ever before[13].

Figure 2: The meeting of two schools of thought.

The HML method, described in the next section, applies the theory of entropy in an indirect way in order to discover structures in noisy sequences. Since estimation of entropy from structured sequences yields incorrect and biased results, HML uses entropy-rate change instead as a learning criterion in the inference of grammars. The method is based on the basic idea that an entropy rate increase occurs when some of the structures in the sequence are destroyed or removed. Therefore, by searching for such a modification that leads to a maximal increase in the sequence entropy rate, one should be able to discover all kinds of hidden structures (patterns) in the sequence and even without any a priori knowledge of what kind of patterns there are and where they are located.

The discovered patterns can be characterized by collecting descriptions of all the modifications made to the sequence during the journey from its low entropy-rate state to its highest possible state. Patterns are discovered by partially or fully removing them. This statistical learning principle seems to be a novelty in the field of grammatical inference and it leads to specific stochastic context free grammars (SCFG), where the
given sequence or a set of sequences (alone) defines the rules and the associated entropy rate changes. These changes indicate the information value of the rule and provide a tool to estimate its production probability.

3. Hybrid Model Learner

Almost 20 years before Shannon’s communication theory was published, G. N. Lewis expressed an idea that “gain in entropy means loss of information” [14]. This general principle is now applied in the HML algorithm. Fig. 3 shows a three dimensional probability space (left) and the corresponding entropy rate hill (right). The latter is a concave function of element probabilities, therefore the gradient at every point of the surface points toward the entropy rate maximum and by searching for such a change in a sequence that causes the largest $\Delta h$ value we have revealed an elementary piece of information, or structure, hidden in the noisy sequence.

In the HML process, the iterative changes made in the sequence are agglomerations of the type $ab \rightarrow c$, where $c$ is a new non-terminal element. If the elementary construction $ab$ is an important substructure of $X$, then $h$ increases. However, this general principle is not always true. If the sequence has an extremely strong pattern (e.g., it is deterministic), like: \{1,2,3,1,2,3,1,2,3,\}, then every agglomeration (during the first steps) may lead to an entropy rate decrease. This is due to the fact that deterministic sequences violate the premises of Shannon entropy leading to erroneous entropy rate estimates.

The idea of agglomeration in connection to sequence compression was discussed by Solomonoff in the 60’s [15, part 4.2]. He writes: “The method to be described begins by considering all possible pairs of symbols. The pair for which the total decrease in “coding cost” is maximum is then assigned a special symbol, and the original sequence is rewritten using this special symbol. At the next stage all pairs of symbols (including the newly defined special symbol) are examined, and the pair for which decrease of coding cost is maximum is assigned a new special symbol”.

This is exactly what HML does – only that the ‘coding cost’ is changed. Further, it appears that the maximal compression of the sequence (optimal coding) doesn’t typically lead to the best grammar for pattern recognition purposes.

Figure 4 illustrates the general structure of the HML process of grammatical inference. The information stream at the input (sounds, genomic sequences, images) is first transformed to the form of discrete sequence whose symbols are from a finite alphabet. In the example given in the next section a 5-point permutation transformation was used to code the audio signal waveform directly to integers (permutation indices, 1 to 120) [4]. The iterative HML process performs a statistical analysis of the entire sequence and selects the element pair $ab$ that leads to maximal increase in entropy rate when replaced with a new element $c$ (the basic idea of Solomonoff is employed). Finally, the obtained rule and the associated entropy rate change are stored.

The obtained new sequence is called the ‘residual’ and the HML process is applied to it repeatedly to produce new, increasingly stochastic residuals. The process may stop when, e.g., the entropy rate doesn’t increase enough or, the residual doesn’t have pairs with frequency higher than two. The process leads to three different representations related to the input sequence: i) a special stochastic context-free grammar (in Chomsky normal form), ii) entropy rate differences of the production rules and, iii) the residual sequence.

The recognition (classification) process may utilize all three of these representations. At this early stage of the work we have only studied a limited range of different variants. We have concentrated on the grammars and entropy rate differences only. Two entropy measures have been used: Shannon entropy and multinomial entropy.

During the recognition phase the rules of the grammar are applied to the unknown sequence in the same order they were extracted from the labeled examples. After every iteration, the residual entropy rate and the sequence length difference are measured. The effect of the grammar $G$ to the unknown sequence is measured by the (total) entropy: $H_G = \Delta L^* h$. Its value in bits is obtained at every iteration $k$. The process is repeated for every model (grammar $G_k$) and the obtained entropy trends $H_G(k)$ are compared to find the winning model. The winning grammar has the largest $H_G$-value, i.e., the grammar with the largest explanatory power.

4. Example: stop consonants /k/, /p/, /t/

The HML process in the form described in the previous section was used to model the bursts of unvoiced stop consonants. Each of the obtained grammars consisted of 600 rules. In this experimental test the HML method with multinomial entropy was applied to the contexts where k-, p-, t- bursts were followed by [uh] appearing in the TIMIT speech corpus. Table 1 summarizes the results.

<table>
<thead>
<tr>
<th>Case</th>
<th>Training</th>
<th>Correct</th>
<th>Perc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[k]-[uh]</td>
<td>33</td>
<td>10/10</td>
<td>100%</td>
</tr>
<tr>
<td>[p]-[uh]</td>
<td>20</td>
<td>9/10</td>
<td>90%</td>
</tr>
<tr>
<td>[t]-[uh]</td>
<td>28</td>
<td>8/8</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 1: Classification test for k-, p-, t-bursts.

The training material consisted of 20-33 cases and tested on 8-10 cases for each consonant, all what TIMIT provides for the [kl]p[t]-[uh] context. Only one error occurs in [p]. Figure 5 illustrates the entropies $H_G(k)$ for the eight cases of [t]-[uh].
The mean of the curves is removed and the difference is integrated in order to produce a smooth output. The x-scale gives the number of rules used divided by 50 (k runs up to 600). It can be seen that in these cases the [t] candidates (the uppermost curves) are clear winners in all cases and the result is stabilized already around k=200. Blue and red curves are for [k] and [p] grammars, respectively.

- how to describe the properties and structures still present in the residual sequence,
- how to present the properties and structures of the unknown input sequence before the agglomeration rules are used and how to combine this with the other measures (classification), and
- what is the role of different preprocessing and different entropy measures in relation to different data types.

Presently the HML method is being extended to utilize parallel grammars inferred from the same sequence. They form different views of the same data thus providing a rich description. Proper combination of these grammars may improve the results further.

6. Conclusions

A new method for constructing specific stochastic (tree-) grammars for discrete sequences was presented. The method is based on agglomeration of adjacent elements that induce the largest entropy-rate change. It was shown that the entropy-rate measure is able to guide grammatical inference to discover and model regularities in sequences.

7. Acknowledgements

The author is grateful to M.Sc. Okko Räsänen for many constructive, illuminating and also critical comments during the final development and preliminary tests of the HML method. His comments as well as those of Ph.D. Toomas Altosaar on this paper are acknowledged.

8. References