

Learning Interpretable Sequence Classifiers through Evolved Regular Expression Patterns

Roberto Canduela Luengo¹, Alberto Fernández-Isabel¹, and
Javier M. Moguerza¹

Rey Juan Carlos University (URJC)
Data Science Laboratory (DSLAB),
Department of Computer Science and Statistics, ETSII
Móstoles, Madrid, Spain
`r.canduela@alumnos.urjc.es`, `alberto.fernandez.isabel@urjc.es`,
`javier.moguerza@urjc.es`

Abstract. Sequence classification methods based on deep neural networks achieve strong predictive performance but often lack interpretability. This limitation is critical in domains where decision transparency is required. This work investigates a preliminary framework for supervised classification of symbolic sequences in which models are expressed directly as Regular Expressions (Regex). Classification is formulated as a pattern-matching problem, and candidate symbolic rules are discovered through a search-based induction process. Preliminary experiments on controlled synthetic datasets suggest that evolving symbolic expressions can capture patterns consistent with the underlying generative structure while producing inherently interpretable classifiers. These findings highlight the potential of symbolic pattern evolution for transparent sequence learning.

Keywords: Sequence classification · Interpretability · Regular expressions · Evolutionary search

1 Introduction

Sequential data appears in a wide range of domains, including bioinformatics, cybersecurity, anomaly detection, and language technologies. In these contexts, inputs consist of ordered symbolic elements whose structure, position, and repetition patterns influence class membership. Although deep neural models such as LSTMs and transformers dominate current solutions, they operate through distributed representations that obscure decision logic. This lack of transparency limits their adoption in regulated and safety-critical environments.

Symbolic learning approaches offer a different paradigm: instead of learning hidden representations, they encode knowledge explicitly through rules. Regular expressions constitute a compact formalism for describing structural properties of sequences, including ordering constraints, optional segments, repetition patterns, and alternative subsequences. This work examines whether supervised

sequence classification can be formulated as the discovery of regex-based decision rules.

Many sequence classification tasks depend not only on symbol frequency but on structural relationships such as subsequences, ordering constraints, and contextual configurations. Symbolic formalisms allow these dependencies to be represented explicitly, and regular expressions offer a hypothesis space that balances expressive capacity with formal tractability, making them suitable for modeling interpretable decision rules.

To the best of our knowledge, prior work on symbolic sequence learning has typically focused on grammar induction, automata inference, or genetic programming for symbolic expressions, often without explicitly restricting the hypothesis class to human-readable regular expressions optimized under a supervised predictive objective.

Thus, the objective of this paper is to assess the feasibility of supervised sequence classification under an explicitly regex-based hypothesis class learned via search-based induction, and to discuss its potential as a transparent alternative to black-box sequence models.

The remainder of this paper is organized as follows. Section 2 reviews related work on sequence classification, highlighting differences between neural, probabilistic, and symbolic approaches. Section 3 describes the proposed regex-based classification framework and the search-based induction perspective underlying the method. Section 4 concludes the paper and outlines directions for future research.

2 Related Work

Sequence classification has been addressed through probabilistic, neural, and symbolic paradigms. Hidden Markov Models (HMMs) [1] model latent state transitions but require structural assumptions, while Conditional Random Fields (CRFs) [2] provide discriminative modeling at the cost of feature engineering. Deep learning approaches, particularly Long Short-Term Memory (LSTM) networks [3] and transformer architectures [4], capture long-range dependencies effectively but rely on distributed representations that are difficult to interpret, motivating interest in transparent alternatives [6].

Symbolic approaches represent knowledge through explicit structures such as automata and regular languages [5]. These models are inherently interpretable, yet typically depend on manually defined rules or fixed structures. The supervised discovery of symbolic sequence classifiers directly expressed as regular expressions through search-based strategies remains comparatively underexplored, which motivates the framework proposed in this paper.

Related lines of research include genetic programming, automata learning, and sequential rule mining, which also explore structured hypothesis spaces for sequences but do not typically restrict the model class to human-readable regex rules optimized under a supervised objective.

3 Regex-Based Classification Framework

Let $E = \{e_1, \dots, e_n\}$ be a discrete domain of symbols and $S = \langle e_{i_1}, \dots, e_{i_m} \rangle$ a labeled sequence with class $c \in \{0, 1\}$. The proposed framework formulates sequence classification as a symbolic pattern recognition problem in which each classifier is represented by a Regular Expression (Regex) acting as an explicit decision rule. Sequences are encoded into linear symbolic strings using a delimiter-based representation that preserves ordering. Given a candidate pattern R , prediction is defined as:

$$\hat{c}(S) = \begin{cases} 1 & \text{if } R \text{ matches the encoded sequence } S, \\ 0 & \text{otherwise.} \end{cases}$$

Under this formulation, learning becomes the task of discovering a pattern R that maximizes classification performance while maintaining structural interpretability through constrained pattern complexity.

Over the symbol domain E , regular expressions provide a compact formalism capable of modeling multiple structural properties of sequences. These include positional constraints at the beginning or end of the sequence, variable-length dependencies through repetition operators, alternative subsequences via logical disjunction, and ordered composition of symbols. Together, these mechanisms allow regex classifiers to capture patterns that depend not only on symbol presence but also on their arrangement, spacing, and contextual positioning.

Manually specifying such patterns is infeasible in most real-world settings. The framework employs a search-based induction process. Candidate regex patterns are represented as structured symbolic expressions generated from a pre-defined regular-expression grammar. Variation operations modify substructures of these expressions while preserving syntactic validity, allowing exploration of alternative symbol combinations, operators, and structural configurations within the hypothesis space.

The search is guided by a supervised performance objective computed over labeled sequences. In addition to predictive performance, the framework allows the incorporation of structural complexity considerations, since excessively long or intricate expressions may reduce interpretability even if they improve training accuracy.

From a modeling perspective, the use of regular expressions defines a hypothesis space in which structural assumptions about sequences are made explicit. This makes it possible to reason about both predictive performance and the form of the decision rule itself. In contrast to models where complexity is distributed across many parameters, here the form of the hypothesis directly reflects the type of dependencies being modeled. This correspondence between model form and decision behavior provides a transparent link between data patterns and decisions, which is particularly valuable in settings where domain constraints or prior knowledge must be considered.

Regex representations are inherently restricted to regular languages and cannot model arbitrary long-range nested dependencies. However, many real-world

sequence classification problems involve structural patterns that are well captured by regular constructs. In such contexts, the gain in interpretability and transparency may outweigh the loss in expressive generality.

Although regex-based classifiers are more transparent than distributed neural models, interpretability is not guaranteed for arbitrarily complex expressions. Very long or highly nested patterns may become difficult to understand, highlighting the need to balance expressiveness and structural simplicity when searching in the symbolic hypothesis space.

4 Conclusions

This paper has investigated the feasibility of learning sequence classifiers expressed as regular-expression patterns discovered through search-based induction. The approach reframes supervised sequence classification as the discovery of symbolic structural rules rather than the optimization of opaque parametric models.

The main contribution lies in intrinsic interpretability, as decision logic is encoded directly in human-readable patterns, enabling inspection, validation, and traceability. While regex representations are limited to regular language expressiveness, many real-world sequence classification problems involve structural patterns that can be effectively captured within this formalism.

Future work will focus on extending symbolic pattern models while preserving transparency, with the aim of reinforcing intrinsic explainability through the supervised extraction of symbolic rules from sequence classifiers. Special attention will be devoted to assessing scalability and generalization across heterogeneous real-world settings. Furthermore, incorporating explanatory visualizations of the proposed workflow could improve the clarity and communicability of the methodology. Altogether, this line of research may help consolidate symbolic pattern evolution as a complementary paradigm for interpretable sequence learning in structure-driven domains.

Acknowledgments

This work has been funded by the Spanish MICIU under the PDI program in the XMIDAS project (PID2021-122640OB-I00).

References

1. Rabiner, L.R.: A tutorial on hidden Markov models. Proc. IEEE (1989)
2. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields. ICML (2001)
3. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation (1997)
4. Vaswani, A. et al.: Attention is all you need. NeurIPS (2017)
5. Kleene, S.C.: Representation of events in nerve nets and finite automata. Princeton (1956)
6. Rudin, C.: Stop explaining black box models. Nature Machine Intelligence (2019)