

Chapter 1

INTRODUCTION TO SYNTACTIC PATTERN RECOGNITION

1.1. SUMMARY

In this chapter we discuss the fundamental idea, system, methods, and applications of syntactic pattern recognition; the reason to use syntactic methods in seismic data. Also we describe the content of each chapter.

1.2. INTRODUCTION

Syntactic pattern recognition has been developed over two decades, received much attention and applied widely to many practical pattern recognition problems, such as (1) English and Chinese character recognition, (2) fingerprint recognition, (3) speech recognition, (4) remote sensing data analysis, (5) biomedical data analysis in chromosome images, carotid pulse waves, EEG signals, . . . , etc., (6) scene analysis, (7) texture analysis, (8) 3-D object recognition, (9) two-dimensional mathematical symbols, (10) spark chamber pictures, (11) chemical structures, (12) geophysical seismic signal analysis, . . . , etc. [3, 6, 13–16, 19, 22–24, 30, 37, 39, 41, 46, 48, 49, 53–55, 58, 59, 62, 64–66, 72, 73, 78–81, 85–89, 92, 96, 98, 100, 112, 113, 116].

In the pattern recognition problems, besides the statistical approach, the structural information that describes the pattern is important, so we can use syntactic methods to recognize the pattern. A pattern can be decomposed into simpler subpatterns, and each simpler subpattern can be

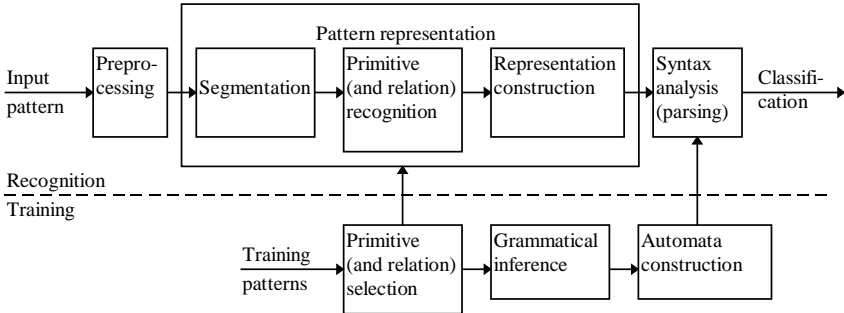


Fig. 1.1. Block diagram of a syntactic pattern recognition system.

decomposed again into even simpler subpatterns, and so on. The simplest subpatterns are called primitives (symbols, terminals). A pattern can be described as a representation, i.e., a string of primitives, a tree, a graph, an array, a matrix, or an attributed string, . . . , etc. [33, 43, 64–66, 68, 91, 110]. We can parse the representation and assign the pattern to its correct class.

A basic block diagram of the syntactic pattern recognition system is shown in Fig. 1.1. The system consists of two major parts: training and recognition. The training part consists of primitive (and relation) selection, grammatical inference, automata construction from the training patterns, and the recognition part consists of preprocessing, segmentation or decomposition, primitive (and relation) recognition, construction of pattern representation, and syntactic parsing analysis for the input testing pattern.

The finite-state grammar, context-free grammar and context-sensitive grammar of the formal language are adopted in the description of 1-D string representation of the pattern [2, 41]. The 1-D string grammars also include programmed grammar, indexed grammar, grammar of picture description language, transition network grammar, operator precedence grammar, pivot grammar, plex grammar, attributed grammar, . . . , etc. [16, 36, 37, 39, 41, 49, 55, 90, 92, 95, 97, 101]. The syntactic parsing analyses include finite-state automata, pushdown automata, top-down parsing, bottom-up parsing, Cocke-Younger-Kasami parsing, Earley's parsing, . . . , etc. [2, 41].

The description power can be extended from 1-D string grammars to high-dimensional pattern grammars for the analysis of 2-D and 3-D patterns. The high-dimensional pattern grammars include tree grammar,

array grammar, web grammar, graph grammar, shape grammar, matrix grammar, . . . , etc. [17, 33, 41, 43, 66, 68, 88, 91, 94, 110]. The syntactic parsing analyses include tree automata, array automata, . . . , etc.

For consideration of substitution, insertion, and deletion errors in the pattern, the automata can be expanded to error-correcting automata to accept the noisy pattern or distorted pattern [1, 64, 65, 84, 101, 102, 106, 115]. The 1-D string grammars and high-dimensional pattern grammars also include stochastic grammars, languages, and the corresponding parsers [29, 40, 44, 86, 101, 105].

The use of pattern recognition has become more and more important in seismic exploration [4, 5, 10, 11, 18-21, 26, 50, 52-68, 99]. However, most of the papers emphasize statistical seismic pattern recognition. Interpreting a large volume of seismic data is a challenging problem. Seismic data in the one-shot seismogram and stacked seismogram may contain some physical and structural information from the response of subsurface. So before interpreting seismic data, it is better to have the techniques to process the seismic data and to improve seismic interpretation. Here using the structural information of seismic data, we propose the important syntactic approach to seismic pattern recognition.

1.3. ORGANIZATION OF THIS BOOK

In Chapter 2, we start to discuss the fundamental theory of formal languages and parsing methods. There are four kinds of basic grammars and languages: finite-state, context-free, context-sensitive, and unrestricted. Finite-state automaton can recognize the finite-state language. Earley's parsing algorithm can recognize the context-free language. Finite-state grammar can be inferred from sample strings. Levenshtein distance is the distance computation between two strings.

In Chapter 3, syntactic pattern recognition techniques are applied to the analysis of 1-D seismic traces to classify Ricker wavelets. Seismic Ricker wavelets have structural information in shape, and each Ricker wavelet can be represented by a string of symbols. To recognize the strings, we use a finite-state automaton to identify each string. The automaton can accept strings having substitution, insertion, and deletion errors of the symbols. There are two attributes, terminal symbol and weight, in each transition of

the automaton. A minimum-cost error-correcting finite-state automaton is proposed to parse the input string.

Two methods of parsing attributed string are proposed. One is the modified error-correcting Earley's parsing in Chapter 4, and the other is a parsing using the match primitive measure (MPM) in Chapter 5.

In Chapter 4, the modified minimum distance error-correcting Earley parsing for an attributed string can handle three types of error. The recognition criterion of the modified Earley's algorithm is "minimum-distance." We discuss the application of the parsing method to the recognition of seismic Ricker wavelets and the recognition of wavelets in real seismic data in Chapter 5.

In Chapter 5, the computation of the match primitive measure between two attributed strings using dynamic programming is proposed. The MPM parsing algorithm for an attributed string can handle three types of error. The MPM parsing algorithm is obtained from the computation between the input string and the string generated by the attributed grammar. The MPM parsing is more efficient than the modified Earley's parsing. The recognition criterion of the MPM parsing algorithm is "maximum-matching". The parsing method is applied to the recognition of seismic Ricker wavelets and the recognition of wavelets in real seismic data.

In Chapter 6, Levenshtein distance computation is applied to detect the candidate bright spot, trace by trace, in the real seismograms. The system for one-dimensional seismic analysis includes a likelihood ratio test, optimal amplitude-dependent encoding, probability of detecting the signal involved in the global and local detection, plus minimum-distance and nearest-neighbor classification rules. The relation between error probability and Levenshtein distance is proposed.

In Chapter 7, tree automaton of syntactic pattern recognition is adopted to recognize 2-D structural seismic patterns. The tree automaton system includes two parts. In the training part of the system, the training seismic patterns of known classes are transformed into their corresponding tree representations. Tree representations can infer tree grammars. Several tree grammars are combined into one unified tree grammar. Tree grammar can generate the error-correcting tree automaton. In the recognition part of the system, each input testing seismogram passes through pre-processing and tree representation of seismic pattern. Each input tree is parsed and recognized into the correct class by the error-correcting tree

automaton. Because of complex variations in the seismic patterns, three kinds of automaton are adopted in the recognition: weighted minimum distance structure preserved error-correcting tree automaton (SPECTA), modified maximum-likelihood SPECTA, and minimum distance generalized error-correcting tree automaton (GECTA). Weighted minimum distance SPECTA and modified maximum-likelihood SPECTA take only substitution errors of the tree structure into consideration. Minimum-distance GECTA takes substitution, deletion, and insertion errors of the tree structure into consideration. The bright spot seismic pattern is shown as the example in the parsing steps. Tree automata could be applied to the recognition of other seismic patterns, such as pinch-out, flat spot, gradual sealevel fall, and gradual sealevel rise patterns. The tree automaton system provides a tool for recognition of seismic patterns, and the recognition results can improve seismic interpretation.

In Chapter 8, we present a hierarchical system to recognize seismic patterns in a seismogram. The seismic patterns are hierarchically decomposed or recognized into single patterns, straight-line patterns or hyperbolic patterns, using syntactic pattern recognition. The Hough transformation technique is used for reconstruction, pattern by pattern. The system of syntactic pattern recognition includes envelope generation, a linking process in the seismogram, segmentation, primitive recognition, grammatical inference, and syntax analysis. The seismic patterns are automatically recognized and reconstructed.