Available online at www.sciencedirect.com
ScienceDirect
Fusion Engineering and Design 83 (2008) 421–424
Fusion Engineering and Design
www.elsevier.com/locate/fusengdes

Structural pattern recognition methods based on string comparison for fusion databases

S. Dormido-Canto [a,*], G. Farias [a], R. Dormido [a], J. Vega [b], J. Sánchez [a], N. Duro [a], H. Vargas [a], G. Rattá [b], A. Pereira [b], A. Portas [b]

[a] *Dpto. Informática y Automática - UNED 28040, Madrid, Spain*
[b] *Asociación EURATOM/CIEMAT para Fusión, 28040, Madrid, Spain*

Available online 4 March 2008

Abstract

Databases for fusion experiments are designed to store several million waveforms. Temporal evolution signals show the same patterns under the same plasma conditions and, therefore, pattern recognition techniques allow the identification of similar plasma behaviours. This article is focused on the comparison of structural pattern recognition methods. A pattern can be composed of simpler sub-patterns, where the most elementary sub-patterns are known as primitives. Selection of primitives is an essential issue in structural pattern recognition methods, because they determine what types of structural components can be constructed. However, it should be noted that there is not a general solution to extract structural features (primitives) from data.

So, four different ways to compute the primitives of plasma waveforms are compared: (1) constant length primitives, (2) adaptive length primitives, (3) concavity method and (4) concavity method for noisy signals. Each method defines a code alphabet and, in this way, the pattern recognition problem is carried out via string comparisons. Results of the four methods with the TJ-II stellarator databases will be discussed.

Keywords: Structural pattern recognition; Data mining; Nuclear fusion; Information retrieval

1. Introduction

Identification problems involving time-series data (or waveforms) constitute a subset of pattern recognition applications that is of particular interest because of the large number of domains that involve such data (for instance, fusion databases). There are some previous works on pattern recognition in fusion databases. Refs. [1,2] are focused on general data mining methods devoted to analysing time series data. However, the goal of our approach is not knowledge extraction but to provide users with an easy tool to perform a first data screening. In this sense, earlier approaches concentrated the efforts in looking for similar full waveforms, i.e. signals covering the full plasma life [3–5]. In another approach, the interest is focused on searching for specific patterns within waveforms [6].

The algorithms used in pattern recognition systems are commonly divided into two tasks, as shown in Fig. 1. The description task transforms data collected from the environment into features (primitives). The classification task arrives at an identification of patterns based on the features provided by the description task.

There is no general solution for extracting structural features from data. The selection of primitives by which the patterns of interest are going to be described depends upon the type of data and the associated application. The features are generally designed making use of the experience and intuition of the designer.

This article summarizes different structural pattern recognition methods and shows specific examples in waveforms of the database of TJ-II stellarator. The difference among these methods is the way in which the primitives are computed. Section 2 describes the main concepts to consider in each technique. Section 3 emphasizes in the application scheme. Finally in Section 4 illustrative examples are shown.

2. Computation of primitives

The selection of primitives is an essential issue because they determine what types of structural components can be constructed.

* Corresponding author. Tel.: +34 91 3987194; fax: +34 91 3987690.
E-mail address: sebas@dia.uned.es (S. Dormido-Canto).

0920-3796/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.fusengdes.2007.11.009

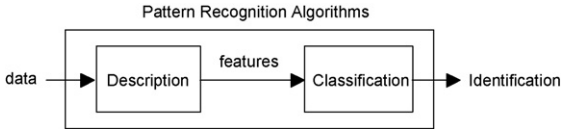

Fig. 1. Tasks in the pattern recognition systems.

We define four ways to compute the primitives of the waveforms: constant length primitives (CLP) [6], adaptive length primitives (ALP), concavity method (CM) and concavity method for noisy signals (CMNS). Each method defines a code alphabet and, in this way, the pattern recognition problem is carried out via string comparisons.

2.1. Constant length primitives

In this method we divide the original signal into segments (all the segments have the same number of samples) where each segment is represented by a straight line. A least square minimization procedure is used to obtain each straight line. Then we encode these segments into a string of primitives. We give a label to each segment and we add the amplitude between the first and the last sample into the primitive. The labelling of the segment $\{(x_i, y_i), (x_j, y_j)\}$ is based on the classification of the slope of the fitted straight line. Our discriminate values, the primitive labels and an illustrative example are depicted in Fig. 2. The classification of the angle gives us all the elementary structural information needed to construct more complex sub-patterns in waveform recognition.

We use five different values (a, c, e, d and z) to represent the classes of the angle. With a larger set of primitive classes we could have expressed more accurately the structure of signal, but the final string would have been more complex. On the other hand these five codes are just enough for the typical plasma evolution analysis. The code e represents a flat part of a signal, codes c and d represent the ascending and descending angle and codes a and z represent the extremely step slopes.

2.2. Adaptive length primitives

As in the CLP method, the original signal is decomposed into L line segments, however in this case, the length of each segment, K, is variable (i.e. the segments do not have a fixed number of samples). A fixed maximun error, E_{max}, is defined for each line segment. This value is a function of the standard deviation of the entire signal. Hence, the ALP methodology is implemented by following these steps:

1. Start of segment (initially $L = 1$).
2. Set K, the number of signal samples to regress (initially $K = 3$).
3. Generate a line regressing.
4. If (fitted error $< E_{max}$), increase K and go to 3, otherwise start a new segment, increase L, and go to 2.

After this processing, the primitives are labelled using the slope thresholds defined for the previous method (CLP).

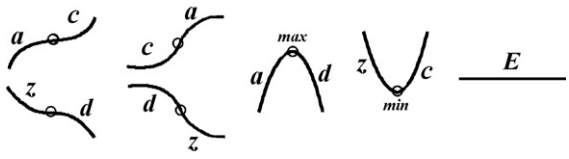

Fig. 3. Stationary points and labels for the classification using concavity methods.

2.3. Concavity method

In this method we detect stationary points of the signal using a simple derivative test. These points belong to one of the six types shown in Fig. 3: maximum, minimum or inflexion point.

Notice that stationary points of inflexion can be easily classified in four types depending on the change in concavity of the curve at that point. Then we give a label to the piece of signal between two consecutive stationary points. The label of each piece of signal $\{(x_i, y_i), (x_j, y_j)\}$ is based on the classification of its concavity. As there are only four possibilities to decide between convex or concave and increase or decrease (shown in Fig. 3); four primitives are sufficient to label any piece of signal. We use four different values (d, z, c, and a) to represent these possible changes of concavities (Fig. 3).

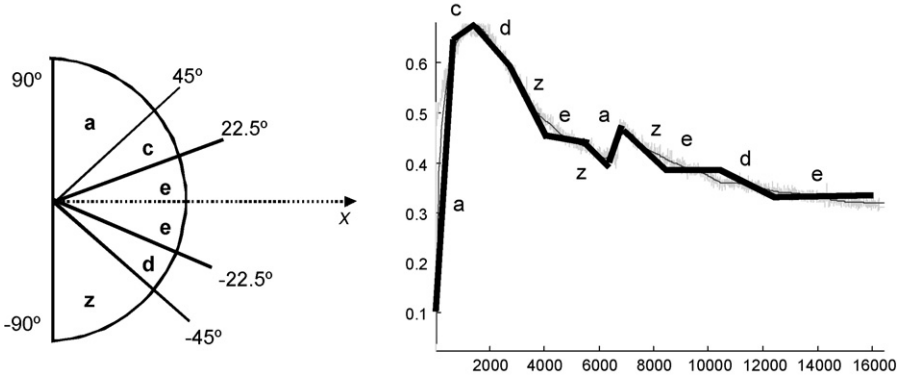

Fig. 2. Discriminates and labels for the classification of the angle of the fitted straight line with an illustrative example.