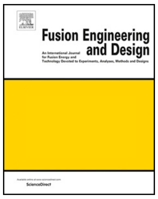
Contents lists available at ScienceDirect

Fusion Engineering and Design

journal homepage: www.elsevier.com/locate/fusengdes



Automatic recognition of plasma relevant events: Implications for ITER

J. Vega[a,*], R. Castro[a], S. Dormido-Canto[b], G.A. Rattá[a], M. Ruiz[c]

[a] Laboratorio Nacional de Fusión – CIEMAT, 28040 Madrid, Spain
[b] Depto. De Informática y Automática, UNED, 28040 Madrid, Spain
[c] Instrumentation and Applied Acoustic Research Group, UPM, Campus Sur, Madrid, Spain

ARTICLE INFO

Keywords:
Automatic recognition of events
Massive databases
ITER
Nuclear fusion
Big data

ABSTRACT

This work makes a proposal about the use of big data techniques for the automatic recognition and classification of plasma relevant events in huge databases of nuclear fusion devices. A relevant event can be any kind of anomaly (or perturbation) in the plasma evolution. This is revealed in the temporal evolution signals as (typically) abrupt variations (for instance in amplitude, noise, or sudden presence/suppression of patterns with periodical structure). A general algorithm based on five steps is presented here for the automatic location and unsupervised classification of plasma events: dataset selection, location of anomalies in individual signals, definition of multi-signal patterns, unsupervised clustering of multi-signal patterns and creation of supervised classifiers. It is important to note that the algorithm implementation is for off-line analysis but supervised classifiers could be implemented under real-time conditions.

1. Introduction

Big data techniques deal with heterogeneous, complex and massive datasets to identify patterns that are hidden inside enormous volumes of data. ITER is expected to acquire more than 1 Tbyte of data per discharge. This amount of data comes from hundreds of thousands of signals acquired in each discharge. Signals can be time/amplitude series, temporal evolution of profiles and video-movies (infra-red and visible cameras). Therefore, the ITER database satisfies the conditions of heterogeneity, complexity and size to use big data techniques for the recognition of hidden patterns.

ITER is a device not focused on basic research on plasma physics. Its aim is to produce high performance plasmas to approach the operation to reactor regimes. Vast amounts of hidden information will remain in the ITER databases and it will be worth to extract as much knowledge as possible about the plasma nature. Due to the large number of signals per discharge and the shot duration (30 min), automatic methods of data analysis will be necessary.

Automatic data analysis methods (ADAMs) can identify relevant temporal segments inside discharges in an automatic way, where the term ‘*relevant*’ means ‘*with interest from some point of view*’ either for physics or for machine control.

ADAMs have a double purpose in the identification of relevant patterns in the databases of nuclear fusion. Firstly, ADAMs make easier the recognition of plasma behaviours by identifying known patterns in experimental signals. Secondly, ADAMs allow the detection of off-normal plasma conditions.

The first purpose is a consequence of the well-known fact that diagnostics produce equal morphological patterns in the signals for reproducible plasma behaviours. It should be noted that the identification of plasma behaviours by means of patterns does not mean that a unique signal defines a pattern. For example, edge localised modes (ELMs) are recognised by synchronous abrupt variations in three different signals: Dα (in the case of deuterium plasmas or Hα for hydrogen plasmas), line integrated electron density and stored diamagnetic energy. In this first purpose, ADAMs can be used to get a better knowledge of the plasma nature as larger databases of known events can be built to obtain better plasma models.

The objective of the second purpose is the potential detection of unknown events that appear on a regular basis. These events can be recognised by the repetition of common patterns that, in principle, are not assigned to known plasma behaviours.

The objective of this article is to make a proposal of a 5-step algorithm to automatically recognise relevant nuclear fusion patterns in massive databases. Section 2 describes several techniques to automatically locate anomalies in signals and Section 3 defines the 5 steps of the algorithm. Section 4 shows a simulation to detect line integrals of plasma emissivity and a particular application of the algorithm presented in this article. Finally, Section 5 is a short discussion.

[*] Corresponding author.
E-mail address: jesus.vega@ciemat.es (J. Vega).

https://doi.org/10.1016/j.fusengdes.2020.111638
Received 9 August 2019; Received in revised form 19 February 2020; Accepted 17 March 2020
Available online 08 May 2020

2. Recognition of plasma relevant events

The high temperature of thermonuclear plasmas compels to measure plasma properties by indirect methods. Plasma diagnostics convert their observations into electrical signals that are digitized. Typically, the conversion of the signals into physics quantities is not a simple calibration factor and, therefore, more or less complex inversion techniques are required.

Visual data analysis of the outputs of diagnostics is the usual way of performing a first screening of discharges. This simple visual analysis allows the identification of times where plasma events happened. In general, a plasma event is shown by unexpected variations of the diagnostic temporal evolution signals. In other words, a plasma in steady state evolves in a quiet way and the diagnostic signals exhibit a smooth evolution. However, the plasma reaction to any kind of perturbation is revealed by means of notable changes in the evolution of signals. In the case of time series, these changes can be in amplitude, noise or presence/suppression of periodical structures. In the case of profiles, plasma events are recognized not only by evident variations in amplitude but also by the generation of hollow profiles, peaked profiles or changes in gradients. For camera diagnostics, video-movies show variations in the emission detected.

Obviously, visual data analysis is not an adequate method to identify abrupt changes in temporal evolution signals 30 min long. As mentioned, plasma events are revealed by abrupt changes that usually take place simultaneously in several signals. Therefore, the automatic location of anomalies in individual signals is a first step to recognize potential plasma relevant events. It should be noted that the more abrupt the change in the plasma evolution the more abrupt is the change of shape in a signal.

Once established that the automatic location of plasma events requires the recognition of abrupt changes in the signals, some techniques for this purpose can be mentioned. The first one is the detection of outliers through a generalized linear regression model. This method is based on the fact that a smooth temporal evolution signal S(t) shows very similar amplitudes between consecutive samples with period τ. This means that a plot in a two-dimensional space whose Y axis is the amplitude at time t, S(t), and whose X axis is the previous sample, S(t – τ), the points are distributed along the diagonal. Samples outside the diagonal are outliers that reveal abrupt changes in the signal. These outliers can be identified with the normal probability plots of residuals (see Fig. 1 as an example).

A second technique to determine outliers in signals is the use of martingales for testing exchangeability [1]. This is a very general technique that requires a single hypothesis in the data stream: samples are independent and identically distributed (*iid*). The samples belong to an unknown probability distribution and they are examined in a sequential way. When the distribution changes (the new distribution is also unknown), the change is detected by testing the exchangeability property of the data. At this moment, an alarm is triggered. It is important to note that the *iid* hypothesis is the usual assumption for the development of machine learning systems.

A third technique for the automatic location of abrupt changes in time series is based on following the temporal evolution of the Fourier components of a signal, which has been explained in this conference [2]. A fourth technique, also presented in this conference [3], uses deep learning methods for the same purposes. Finally, the Universal Multi-Event Locator (UMEL) technique [4] can be used for automatic event location in waveforms and video-movies. Of course, there are many other ways of detecting anomalies in the signals.

3. Algorithm for off-line automatic recognition of plasma relevant events

This section summarises the 5-step algorithm for the Automatic Detection and Unsupervised Classification (ADUC5) of plasma events. It should be emphasised that the software codes that implement the several steps have to be executed in a sequential and unattended way to ensure automatic recognition.

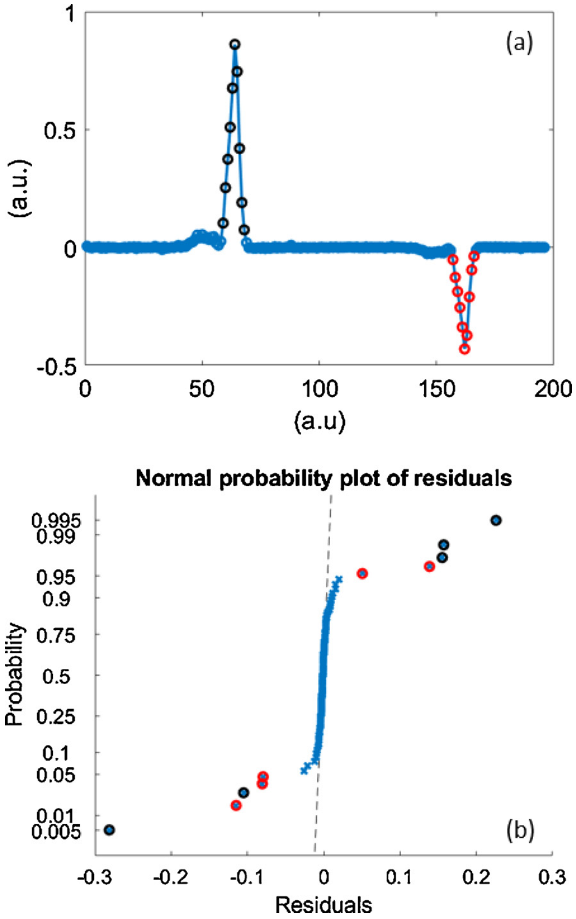


Fig. 1. (a) Abrupt peaks in the temporal evolution of signals. (b) The peaks are detected with the normal probability plots of residuals.

3.1. *Definition of a dataset of signals and a range of discharges*

Given a large database of signals and discharges, this step is used to select a large enough dataset of N_S signals corresponding to a large enough set of N_D discharges. The use of a large N_S allows finding all the signals related to a specific plasma event. In the same way, a large N_D increases the statistical relevance of the results.

3.2. *Determination of times in each discharge where individual signals show anomalies*

Given the selection of data and discharges of Section 3.1, the present step of ADUC5 determines the times t_A where anomalies are detected in the several signals. Fig. 2 is an example in which the anomalies can be grouped in 4 potential plasma events at 4 different times in a single discharge.

3.3. *Definition of multi-signal patterns (MSPs)*

It is important to mention that visual data analysis is useful because it allows identifying plasma behaviours by recognising structural shapes in the signals. So, the contribution of individual signals to detect a plasma event is not a simple amplitude at a given time but the signal shape determined by the samples around the anomaly time t_A (Fig. 3). Therefore, it is necessary to define a time interval around the anomalies

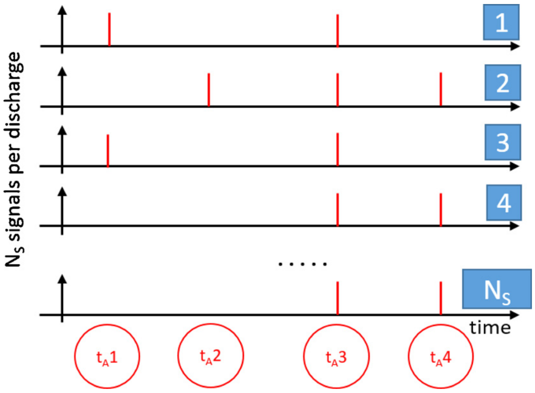


Fig. 2. Vertical lines represent the times when anomalies have been detected in individual signals. It should be noted that the potential plasma event at time t_A2 is only recognized by one anomaly in signal 2. However, the potential plasma event at time t_A3 is detected by anomalies in all signals.

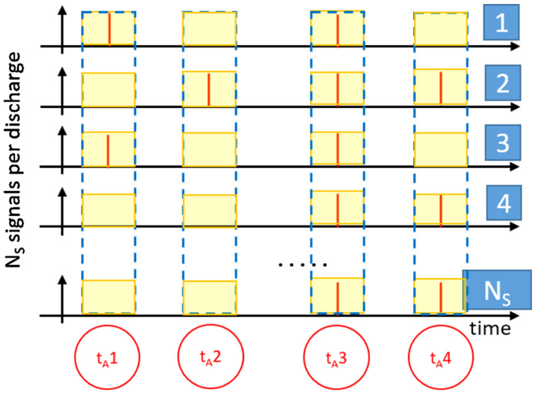


Fig. 3. An interval $t1 < t_A < t2$ is defined around each potential plasma event in order to form multi-signal patterns. The feature vector that represents a MSP is made up of all samples in the yellow rectangles.

$(t_1 < t_A < t_2)$ in such a way that multi-signal patterns are formed. A multi-signal pattern is a feature vector made up of all morphological patterns of the N_S signals within a common time interval $[t_1, t_2]$. By assuming that all signals are sampled with the same period τ, all morphological patterns of individual signals will have M samples. Therefore, $MSP(t_A) \in \mathbb{R}^{M \cdot N_S}$ and

$$\begin{aligned} MSP(t_A) = (&S_1(t1), S_1(t1+\tau), ..., S_1(t2),\\ &S_2(t1), ..., S_2(t2),\\ &...\\ &S_{N_S}(t1), ..., S_{N_S}(t2)) \end{aligned}$$

where $S_j(t)$, $j = 1, ..., N_S$ are the signals that have been chosen in step 1.

It should be noted that a multi-signal pattern contains also the samples of signals that have not shown anomalies in a potential plasma event. On the other hand, a criterion to define the bounds t_1 and t_2 corresponding to the common time interval $[t_1, t_2]$ is necessary. To this end, all MSPs in all discharges of the dataset have to be taken into account. The reason for this resides in the fact that all multi-signal patterns in all discharges must have the same dimensionality ($M{\cdot}N_S$) to carry out the unsupervised clustering of step 4.

3.4. Unsupervised clustering of multi-signal patterns

So far, each potential plasma event in the set of N_D discharges is represented by an MSP (i.e. a feature vector of dimension $M{\cdot}N_S$). However, several questions arise: how many MSPs really represent plasma events? How many different plasma events are present? How many of the plasma events are recognised as known plasma behaviours? What do the rest of MSPs mean?

The answers to the above questions are dealt with step 4 of the ADUC5 algorithm. The objective of step 4 is to group all the MSPs found in step 3 into a number of sensible clusters in an unsupervised way [5–9]. The grouping of the MSPs into clusters provides a classification of the potential events. The different clusters can be labelled with simple tags (lest's say, class A, class B and so on). However, the challenge is to identify each cluster with a physical behaviour of the plasma. Of course, this identification work is to be done by experts and not in an unattended way.

Clusters that are identified with physical behaviours can be used to increase the statistical relevance of the data analysis as it was pointed out in the introduction. Clusters that are not identified with physical behaviours but show statistical weight (lots of MSPs in the cluster) suggest the presence of potential plasma behaviours not recognised so far. In general, they represent off-normal plasma behaviours. Finally, clusters without statistical weight (very few MSPs per cluster) can be considered outliers.

3.5. Development of supervised classifiers with the classes of step 4

Step 4 allows splitting the training MSPs into several well-defined classes. The resulting classification can be used as training dataset of a supervised classifier. This supervised multi-class classifier can be implemented to provide together with each prediction a measure of its reliability. The reliability measure can be a probability, an interval of probability or values of confidence and credibility. These reliable classifiers will allow analysing the robustness of the ADUC5 algorithm. High reliability in the classification of new MSPs with different supervised classifiers will mean a high confidence in the results.

On the other hand, it is important to note that these supervised classifiers can be implemented under real-time conditions, thereby allowing the recognition of behaviours during the execution of discharges. Methodologies able to implement these capabilities were presented in this conference [10].

4. Example of application of ADUC5

This section shows an example of the ADUC5 steps to recognise events. In particular, low order rotating magnetohydrodynamic (MHD) modes have been considered. A continuous sequence of MHD rotating modes haven been simulated according to the following set-up: firstly, an $m = 1/n = 2$ MHD mode that rotates during a time interval t_R; secondly, an $m = 3/n = 2$ mode also rotating during t_R; thirdly, an $m = 4/n = 2$ mode that rotates during a same interval t_R. Therefore, these three sequential rotating modes appear in a periodic way with a period $3{*}t_R$. The objective of the ADUC5 test is to recognise when the transitions from $m = 1/n = 2$ to $m = 3/n = 2$, from $m = 3/n = 2$ to $m = 4/n = 2$ and from $m = 4/n = 2$ to $m = 1/n = 2$ take place.

This test needs the simulation of a diagnostic to acquire data. Fig. 4 shows cross-sections of the plasma emissivity with the MHD modes mentioned above. A tomographic diagnostic is simulated to get integrated measurements of the plasma emission along lines of sight (LOS). The LOS are grouped to form projections where each projection is defined by the set of LOS that cover the whole plasma from the same poloidal angle δ. Fig. 5 shows an array of detectors with a common collimation slit that form a projection. The detectors obtain line integrals of the plasma emission.

The emission intensity detected by each line of sight at time t is the line integral

$$I(\delta, d, t) = \int_{L(\delta,d,t)} E(x, y, t)\, dl$$

where $E(x, y, t)$ is the plasma emission and $L(\delta, d, t)$ is the line of sight corresponding to detector d in projection δ.

Three detector arrays have been chosen for the present simulations.

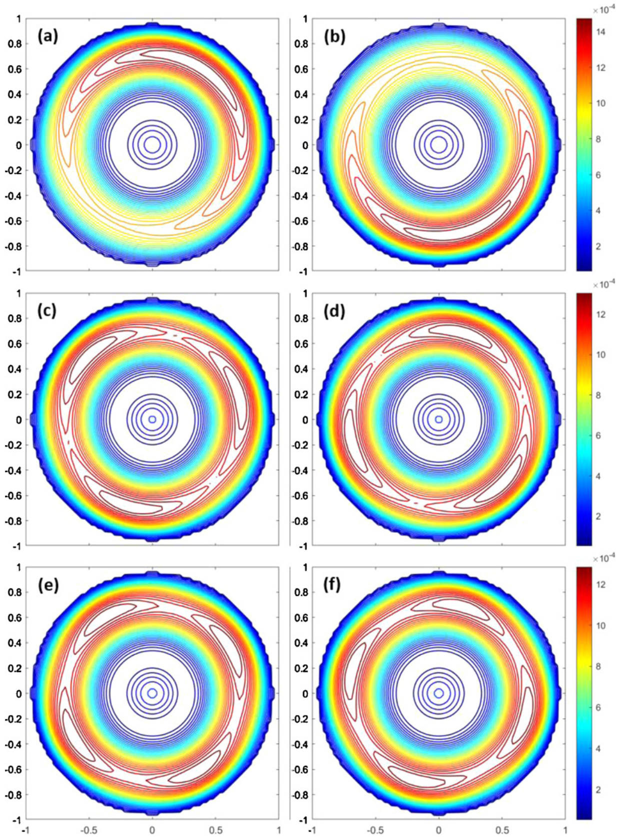


Fig. 4. Simulation of two-dimensional spatial distributions of plasma emission through a cross-section. Plots (a) and (b) show the m = 1/n = 2 mode at different poloidal angle as a consequence of the rotation. Plots (c) and (d) represent m = 3/n = 2 rotating modes and plots (e) and (f) correspond to m = 4/n = 2 rotating modes.

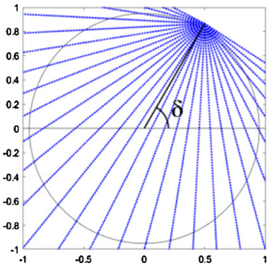


Fig. 5. Projection geometry in a plasma cross-section. The black circle represents the plasma limit and the blue lines are the LOS (all of them share a collimation slit). Radial coordinates in the X and Y axes are in arbitrary units.

All of them have 30 LOS and are located at δ = 0°, δ = 60° and δ = 90° respectively. Fig. 6 is an example of the emission detected by the 90 simulated detectors at the same time instant. In other words, the top, middle and bottom plots in Fig. 6 are respectively the projections

$$P(90^{o}, t) = (I(90^{o}, 1, t), \ldots, I(90^{o}, 30, t))$$
$$P(60^{o}, t) = (I(60^{o}, 1, t), \ldots, I(60^{o}, 30, t))$$
$$P(0^{o}, t) = (I(0^{o}, 1, t), \ldots, I(0^{o}, 30, t))$$

In connection to ADUC5, the temporal evolution of the respective projections are the base signals (step 1 of ADUC5) to look for anomalies in this simulation. For example, Fig. 7 represents the temporal evolution of the δ = 90° projection during a time interval of 3*t_R. The modes evolve as described in the first paragraph of this section.

Taking into account the multi-dimensional nature of the projections to look for anomalies in their temporal evolution, the normalised dot product criterion is applied:

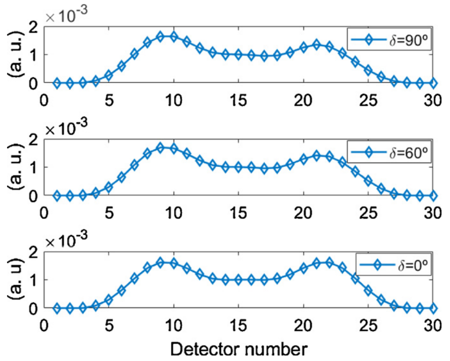


Fig. 6. Line integrals (in a. u.) of the plasma emission detected by the three detector arrays: δ = 90º, δ = 60º and δ = 0º respectively. The X axis represents the number of the detector in the array (clockwise numbered according to Fig. 5).

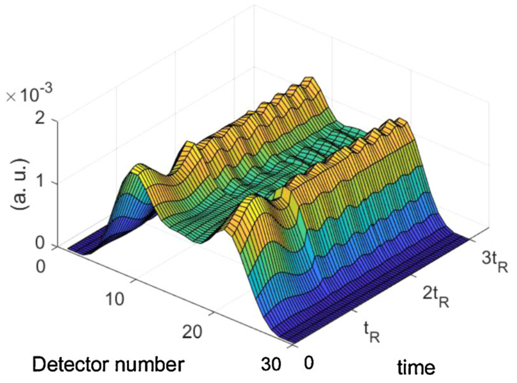


Fig. 7. $P(90^{o}, t = 0, \tau, 2\tau, \ldots, t_R, \ldots, 2t_R, \ldots, 3t_R)$ where τ is the sampling period between projections.

$$\cos\alpha = \frac{|\mathbf{u}\cdot\mathbf{v}|}{||\,\mathbf{u}\,||\cdot||\,\mathbf{v}\,||}, \text{ where } ||.\ || \text{ is the Euclidean norm}$$

According to this, the similarity between two consecutive projections $\mathbf{u} = P(\delta, t - \tau)$ and $\mathbf{v} = P(\delta, t)$ from the same δ angle is maximum if **u** and **v** are parallel and minimum if they are perpendicular (no similarity at all).

By assuming that the similarity between consecutive projections follows a Gaussian distribution, an anomaly will be recognised when the similarity corresponding to two consecutive projections will be outside the interval $\mu \pm 3\sigma$, where μ and σ are, respectively, the mean value and standard deviation of the Gaussian distribution of the similarity.

By applying this criterion to complete step 2 of ADUC5, the temporal evolution of the three projections show simultaneous anomalies at times $t_A = k{\cdot}t_R$, $k = 1, 2, \ldots$, that are precisely the times when the rotating modes change.

At this points, the MSPs of the third step of ADUC5 have to be determined. Following the reasoning of Section 3.3, the MSPs around the anomaly times are:

$$MSP(t_A) = (P(90^{o}, t_A - \tau), P(90^{o}, t_A), \\ P(60^{o}, t_A - \tau), P(60^{o}, t_A), \\ P(0^{o}, t_A - \tau), P(0^{o}, t_A)), \; t_A = k{\cdot}t_R, \; k = 1, 2, \ldots$$

Step 4 of ADUC5 implies now the unsupervised clustering of the previous MSPs. To do this, an agglomerative hierarchical clustering can be carried out by means of a dendrogram. Fig. 8 shows the arrangement of the clusters obtained with the MSPs. It is important to note that MSPs 1, 4, 7 and 10 are grouped together and these patterns represent the

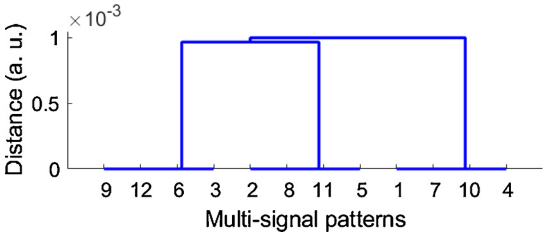


Fig. 8. Hierarchical clustering using a dendrogram.

transitions from m = 2/n = 2 to m = 3/n = 2. In the same way, MSPs 2, 5, 8 and 11 are part of the same cluster and they are showing transitions from m = 3/n = 2 to m = 4/n = 2. Fig. 8 also groups in a single cluster the MSPs corresponding to the transitions from m = 4/n = 2 to m = 1/n = 2 (i.e. MSPs 3, 6, 9 and 12).

From this simulation, two facts are clear from the data analysis of ADUC5, even without knowing the initial set-up of rotating modes. Firstly, the third step of ADUC5 allows putting the focus on specific time instants that recognise anomalies in the temporal evolution of the signals. Secondly, the fourth step identifies three different patterns that appear in a periodic way. These results are important to centre the attention of the data analyst on the anomaly times. In addition, the recurrent observation of repeated patterns gives an important clue about the existence of periodical behaviours. Therefore, looking in detail at different signals around the anomaly times may help in the recognition of changing rotating modes. If so, step 5 of ADUC5 will develop supervised classifiers to identify specific MHD mode transitions whenever the explicit patterns are found. If the patterns cannot be associated to a specific physics behaviour, at least, the unsupervised clustering will allow recognising known patterns although their physics behaviour is not clear.

5. Discussion

The ADUC5 algorithm provides not only the capability of recognizing plasma events but also the possibility of identifying the signals that are more relevant to describe each plasma behaviour. To do this, several executions of the algorithm can be carried out by choosing different datasets of signals in step 1.

Typically, the implementation of ADUC5 requires high performance computing. By assuming 200 sampling times per MSP and $N_S = 100$ signals (for instance 95 time series, 3 profiles 120 points each and 2 video-movies 500 × 300 pixels per frame and 2 bytes per pixel), the amount of required memory is 120 Mbytes/MSP. Now, let's assume 1 relevant event/10 s, the unsupervised classification process requires 720 Mbytes/minute per shot. Thinking of ITER shots (30 min long), this implies 21 Gbytes of memory per shot. By considering a set of $N_D = 500$ discharges, the total memory amount to solve the unsupervised clustering is 10 Tbytes.

The implementation of the ADUC5 algorithm will generate a lot of advanced software codes to locate anomalies with different methods and to classify patterns in both an unsupervised way and a supervised way. The generated codes will be general enough to be used with different signals and different discharges again and again. Therefore, these machine learning codes should be shared in a distributed computed environment, fully accessible from an interactive environment by means of the corresponding authentication and authorization system. Similar tools exist for desktop environments for the Java and Python languages. Weka [11] provides a KnowledgeFlow tool that helps users to apply the different algorithms using a user-friendly Graphical User Interface (GUI). Orange [12] is a GUI for Python data-handling and machine learning algorithms. Both tools provide a number of routines for programmers that can be used by programmers or accessed using a more user- friendly interface. Unfortunately, both tools are designed and implemented for a desktop environment. This fact limits their applicability to relatively small sets of data (as compared to the present large datasets requirements). Also, the algorithms of both tools are generic and are not custom-tailored to the present field. To overcome these problems, a contribution to this conference [13] presented a graphic, data-flow oriented approach able to deal with massive databases.

Finally, it is important to emphasise that the automatic recognition of physics behaviours is only possible if the unsupervised clusters have been labelled by experts. If this categorisation is not possible, the automatic analysis will locate similar patterns in several discharges and temporal locations, but without assigning a physical meaning. This assignation has to be carried out only by specialists.

CRediT authorship contribution statement

J. Vega: Methodology, Investigation. **R. Castro:** Software. **S. Dormido-Canto:** Formal analysis. **G.A. Rattá:** Software. **M. Ruiz:** Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was partially funded by the Spanish Ministry of Economy and Competitiveness under the Projects No. ENE2015-64914-C3-1-R, ENE2015-64914-C3-2-R and ENE2015-64914-C3-3-R.

References

[1] S. Ho, H. Wechsler, A martingale framework for detecting changes in data streams by testing exchangeability, IEEE Trans. Pattern Anal. Mach. Intell. 32 (12) (2010) 2113–2127.
[2] R. Castro, J. Vega, Smart decimation method for fusion research data, 12th IAEA Technical Meeting on Control, Data Acquisition and Remote Participation for Fusion Research. May 13–17, Daejeon (Republic of Korea), 2019.
[3] G. Farias, E. Fábregas, S. Dormido-Canto, J. Vega, S. Vergara, Automatic recognition of anomalous patterns in discharges by recurrent neural networks, 12th IAEA Technical Meeting on Control, Data Acquisition and Remote Participation for Fusion Research. May 13–17, Daejeon (Republic of Korea), 2019.
[4] J. Vega, A. Murari, S. González, A universal support vector machines based method for automatic event location in waveforms and video-movies: applications to massive nuclear fusion databases, Rev. Sci. Instrum. 81 (2010) 11 023505.
[5] A. Mur, R. Dormido, J. Vega, N. Duro, S. Dormido-Canto, Unsupervised event characterization and detection in multichannel signals: an EEG application, Sensors 16 (590) (2016) 14.
[6] A. Mur, R. Dormido, J. Vega, S. Dormido-Canto, N. Duro, Unsupervised event detection and classification of multichannel signals, Expert Syst. Appl. 54 (2016) 294–303.
[7] K. Mehrotra, C. MohanHua, H. Huang, Clustering-Based Anomaly Detection Approaches, Springer, 2017.
[8] Kai Peng, et al., Balanced iterative reducing and clustering using hierarchies with principal component analysis (PBirch) for intrusion detection over Big data in mobile cloud environment, International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage, Springer, Cham, 2018.
[9] Mingrui Zhang, Use density-based spatial clustering of applications with noise (DBSCAN) algorithm to identify galaxy cluster members, IOP Conf. Ser. Earth Environ. Sci. 252 (4) (2019) IOP Publishing.
[10] M. Astrain, M. Ruiz, S. Esquembri, A. Carpeño, E. Barrera, J. Vega, Methodology to standardize the development of FPGA-based intelligent DAQ and processing systems on heterogeneous platforms using OpenCL, 12th IAEA Technical Meeting on Control, Data Acquisition and Remote Participation for Fusion Research. May 13–17, Daejeon (Republ. of Korea), 2019.
[11] Eibe Frank, Mark A. Hall, Ian H. Witten, The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", fourth edition, Morgan Kaufmann, 2016.
[12] J. Demsar, T. Curk, et al., Orange: data mining toolbox in Python, J. Mach. Learn. Res. 14 (August) (2013) 2349–2353.
[13] F. Esquembre, S. Dormido-Canto, J. Vega, G. Farias, J. Chacón, E. Fábregas, Graphic interactive environment for the design of remote processing, analysis and visualization of fusion data, 12th IAEA Technical Meeting on Control, Data Acquisition and Remote Participation for Fusion Research. May 13–17, Daejeon (Republic of Korea), 2019.