EFDA–JET–CP(09)04/05

J. Vega, A. Murari, G.A. Rattá, S. González, S. Dormido-Canto
and JET EFDA contributors

Progress on Statistical Learning Systems as Data Mining Tools for the Creation of Automatic Databases in Fusion Environments

Progress on Statistical Learning Systems as Data Mining Tools for the Creation of Automatic Databases in Fusion Environments

J. Vega[1], A. Murari[2], G. A. Rattá[1], S. González[1], S. Dormido-Canto[3]
and JET EFDA contributors*

JET-EFDA, Culham Science Centre, OX14 3DB, Abingdon, UK

[1]Asociación EURATOM/CIEMAT para Fusión. Avda. Complutense, 22. 28040 Madrid. Spain
[2]Associazione EURATOM-ENEA per la Fusione, Consorzio RFX, 4-35127 Padova, Italy
[3]Dpto. Informática y Automática - UNED, Madrid, Spain
** See annex of F. Romanelli et al, "Overview of JET Results",*
(Proc. 22nd IAEA Fusion Energy Conference, Geneva, Switzerland (2008)).

Preprint of Paper to be submitted for publication in Proceedings of the
7th Technical Meeting on Control, Data Acquisition and Remote Participation for Fusion Research,
Aix-en-Provence, France.
(15th June 2009 - 19th June 2009)

ABSTRACT.

Nowadays, processing all information of a fusion database is a much more important issue than acquiring more data. Although typically fusion devices produce tens of thousands of discharges, specialized databases for physics studies are normally limited to a few tens of shots. This is due to the fact that these databases are almost always generated manually, which is a very time consuming and unreliable activity. The development of automatic methods to create specialized databases ensures first, the reduction of human efforts to identify and locate physical events, second, the standardization of criteria (reducing the vulnerability to human errors) and, third, the improvement of statistical relevance. Classification and regression techniques have been used for these purposes. The objective has been the automatic recognition of physical events (that can appear in a random and/or infrequent way) in waveforms and video-movies. Results are shown for the JET database.

1. INTRODUCTION

This article reviews recent developments in classification and regression techniques for the automatic generation of specialized databases of physical events. These methods have been used to determine the precise time instants in which events happen.

Physical systems can evolve in such a way that different states can occur in different time intervals. The exact recognition of the system state at each time instant can be essential not only for the interpretation of the physics but also for control purposes. Automatic classification systems can be used to follow the temporal evolution of physical systems and to determine the transition times between different states.

Thermonuclear plasmas show two different confinement regimes: Low mode (L-mode) and High mode (H-mode). Section 3 is devoted to the automatic determination of transition times between these confinement regimes in JET. Results of a previous work are reviewed and new requirements in terms of computing performance are also emphasized.

Long pulse fusion devices such as W7X or ITER will have databases with an extremely large number of signals and very long records. Off-line analysis under these conditions requires automatic mechanisms to search for relevant data.

With regard to this, two facts should be highlighted. On the one hand, some plasma events can appear very frequently, but on non-periodic basis, and they have to be located with accuracy for physical studies. Nowadays, this happens for example with Edge Localized Modes (ELMs). On the other hand, a first screening of very long records should provide signal intervals with potential interesting events in order to discard irrelevant time segments. Both of these requirements can be tackled with regression estimations based on SVM. This is an innovative method that is summarized in section 4. New results are presented about, first, the automatic exact location of ELMs in the JET database and, second, about the automatic identification of interesting time intervals in video-movies.

All computations have been performed with Matlab and, in particular, the SVM implementation has been 'The Spider' software, which is included in public licensed environments for Matlab [1].

2. SPECIALIZED DATABASES IN FUSION

Most of the diagnostics in fusion are devoted to following the temporal evolution of quantities. The analysis of physical events (for example sawteeth activity, ELMs, internal transport barriers or disruptions) requires the development of specialized large databases around their time instants. This is an essential step to provide analyses of the physical phenomena with adequate statistical significance.

However, the fusion community suffers from a chronic data analysis issue: only a very limited part of the collected information is properly analyzed. The reason for this is the way in which specialized databases are generated. Typically, the exact temporal location of events is accomplished through the visual inspection of the data. This manual process is very time consuming, mainly for physical phenomena that occur at random or infrequently.

In general, events are recognized by representative signatures (patterns) in the signals. These footprints are characterized by local information either in the time domain or in the frequency domain or in both. With the typical manual procedures, determining accurate time instants can be a difficult task due to several factors. First, the event recognition can require the identification of patterns in noisy data. Second, the signature can present a not perfectly-defined temporal location. Third, specific signal pre-processing can be needed for the temporal location of patterns that are only apparent in the frequency domain (for instance, some MHD phenomena). Fourth, image diagnostics (infrared and visible cameras) are becoming very popular systems, but under long pulse conditions, the location of events through video-movies visualization by human operators is completely unfeasible. Fifth, some physical behaviours can only be detected by means of the simultaneous presence of typical structural forms in different signals (both waveforms and video-films). Sixth, an important issue in the last case can be the diverse sampling frequencies of several signals to determine an individual time instant.

Therefore, for next fusion devices (W7X or ITER) and also for present machines, automatic methods of data analysis are required. In this sense, an initial step can be the automatic location of physical events and the automatic identification of signal intervals with potential relevant data.

The automatic character of these techniques provides clear benefits. On the one hand, practically all the information inside the databases can be used, thereby improving the statistical relevance. On the other hand, the automatic procedures allow a drastic reduction of human efforts in the identification and location of physical behaviours. Finally, a very important goal can be achieved: the search of events can be standardized. This implies that the recognition criteria can be established and implemented by computational means. The consequence is a more objective search mechanisms and less vulnerability to human judgements (missing occurrences, subjective assessments or location errors).

3. CLASSIFICATION TECHNIQUES: L-H TRANSITIONS IN JET

In the general formulation of the classification problem, an input sample (or feature vector) $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ needs to be classified to one (and only one) of J groups (or classes) $C_1, C_2, \ldots, C_J$. To this end, a decision rule is used [2].

A first work on the use of classification systems to determine L-H transition times in JET is [3]. In this work, a set of 50 discharges as training/test datasets between Pulse No's: 52211 and 62723 is considered. The transition times have been established by experts with a high degree of confidence [4]. The feature vectors are made up of six components. To characterize the L to H transition, six waveforms are used: the coordinates of the X-point, the beta normalized with respect to the diamagnetic energy, the total heating power, the magneto-hydrodynamic energy and the axial toroidal magnetic field at a $\Psi = 0.8$ surface. For the H to L transition the last three signals are replaced by the toroidal magnetic field, the safety factor and the time derivative of the diamagnetic energy. The training process has been performed with 6600 and 6000 feature vectors (at sampling periods of 10 ms) for the L-H and H-L transitions respectively.

For a proper transition time evaluation, it is important that the classification system reaches high success rates in discriminating between the two confinement regimes. In this respect, in the reference [3] a combination of classifiers has been shown to significantly increase the success rates obtained with single classifiers [5]. In particular, a combination of a SVM classifier with a Bayes classifier has been developed in [3] as explained in the following.

3.1 SVM CLASSIFIER

The SVM classifier calculates a separating hyper-plane as a decision rule [2] to discriminate between the two classes of confinement regimes: C_L (L-mode) and C_H (H-mode). Given a feature vector at a time instant, the classifier estimates not only the class of **x** but also the distance to the separating hyper-plane. The larger the distance the more reliable is the classification. Then, it is possible to assign a probability to the classification result through a sigmoid function. Therefore, the probability with the SVM classifier that the plasma is in L mode is:

$$P_{SL} = \frac{1}{1 + \exp\left[-kD(\mathbf{x})\right]}$$

where $|D(\mathbf{x})|$ is the distance to the hyperplane and the sign of $D(\mathbf{x})$ distinguishes between L mode or H mode (+1, −1 respectively). The coefficient k is determined empirically. Obviously, the probability that the plasma is in the H regime is $P_{SH} = 1 - P_{SL}$.

3.2 BAYES CLASSIFIER

Given the classification task of **J** classes and an unknown pattern **x**, the Bayes rule states

$$P\,(C_j|\mathbf{x}) = \frac{P\,(\mathbf{x}|C_j)\,P\,(C_j)}{p\,(\mathbf{x})}$$

In words, the Bayes formula gives the *posterior probability* that the unknown pattern belongs to the respective class C_j, given that the corresponding feature vector takes the value **x**. The probability distribution function (pdf) $p\,(\mathbf{x}|C_i\,)$ is the likelihood function of C_j with respect to **x**. $P\,(C_i\,)$ is the a *priori probability* of class C_j and p (**x**) is the pdf of **x** given by

$$p\,(\mathbf{x}) = \sum_{j=1}^{j} p\,(\mathbf{x}|C_j)P\,(C_j)$$

In order to apply the Bayes rule, the likelihood and the prior probability of each class must be known. The likelihood can be estimated via the non-parametric Parzen window estimator [6]. The particular details on the specific implementation to the JET database are discussed in detail in [3]. Concerning the prior probability and taking into account that the plasma is either in L mode or H mode, it is possible to assume a value of $P\,(C_i) = 0.5$ for both cases.

3.3 COMBINED CLASSIFIER

The fusion of the previous classifiers is carried out by means of a fuzzy aggregation operator [7]. In particular, we have chosen the Einstein sum that for the case of the H mode is given by:

$$S_H|\mathbf{x}) = \frac{P_{SH} + P\,(C_H|\mathbf{x})}{1 + P_{SH}\,P\,(C_H|\mathbf{x})}$$

The final decision of this hybrid classifier is based on the value of S_H: if $S_H \geq 0.5$, the plasma is in H mode. Otherwise, the plasma is in L mode.

The combination of classifiers gives success rates of 99.22% for L to H transition and 96.31% for the H to L case.

Figure 1 shows an example of the time estimate that corresponds to a transition from the L to H regime in JET. Feature vectors are taken with a period of 10ms. The vertical green line (at time instant 13.39s) represents the transition time determined by the experts. The horizontal straight line with value 0.5 shows the threshold value to discern between L and H mode. The continuous line is the temporal evolution of the Einstein sum operator. The intersection of this line with the horizontal straight line defines the transition time. Points at values +1/−1 indicate plasma in L and H mode respectively. The transition is estimated to occur at 13.41 ± 0.01s by the combined classifier. The difference between the time determined by experts and the time deduced by the combined classifier is (in absolute value) 0.02s. Similar results are shown in [3] for the H-L transition.

3.4 REMARKS ON THE CLASSIFIER

With a reduced number of discharges, a classifier with enough generalization capability can be developed. However, the temporal resolution of the transition is limited by the sampling period of the feature vectors (10ms in the present case). In order to increase the resolution, let's say up to 1ms, the training process would need to multiply the number of input samples by a factor of 10. Computational times for the training process rise exponentially with the number of feature vectors. For example, to obtain the SVM model, the optimization problem to solve can take 60-90 minutes with 6000 samples in a PC with Matlab. In case of 60000 feature vectors, computing times can be of the order of 70 hours in the same PC. This means that parallelization techniques can be extremely beneficial not only to increase the resolution times but also to augment the number of training discharges to achieve higher generalization capabilities [8].

4. REGRESSION TECHNIQUES: AUTOMATIC LOCATION OF EVENTS IN JET

Reference [9] describes an innovative technique (UMEL, Universal Multi-Event Locator) to automatically locate both singular events and slices of potential interest in any kind of signals (waveforms, images or, in general, multivariate signals with arbitrary number of dimensions). UMEL is based on SVM regressions. Let us consider S training samples $(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_S,y_S)$, $(\mathbf{x}_i \in \mathbb{R}^n$ and $y_i = f(\mathbf{x}_i)$ where $f:\mathbb{R}^n \rightarrow \mathbb{R})$. The regression function is given by [2]

$$f^*(\mathbf{x}) = \sum_{k=1}^{S} \gamma_k^* H(\mathbf{x}_k, \mathbf{x}) \tag{1}$$

The parameters γ_k^*, $k = 1,\ldots,S$ are determined using the solution of the following quadratic optimization problem:

$$\gamma_k^* = \alpha_k^* - \beta_k^*, \; k = 1,\ldots, S$$

where the parameters α_k^*, β_k^*, $k = 1,\ldots,S$are determined by maximizing the functional

$$Q(\alpha, \beta) = -e \sum_{k=1}^{S} (\alpha_k + \beta_k) + \sum_{k=1}^{S} y_k (\alpha_k - \beta_k) - \frac{1}{2} \sum_{k,l=1}^{S} (\alpha_k - \beta_k)(\alpha_l - \beta_l) \; H(\mathbf{x}_k - \mathbf{x}_l)$$

subject to the constraints

$$\sum_{k=1}^{S} \alpha_k = \sum_{k=1}^{S} \beta_k, \; 0 \quad \alpha_k \quad \frac{C}{S}, 0 \quad \beta_k \quad \frac{C}{S}, \; k = 1,\ldots S$$

given the training data $(\mathbf{x}_k, y_k)$, $k = 1,\ldots S$, an inner product kernel H, an insensitive zone e, and a regularization parameter C.

Only a subset of the parameters γ_k^* in (1) Reference source not found. is nonzero. The data points x_k associated with the nonzero γ_k^* are called support vectors.

Different combinations of kernels, e-values and regularization parameter allow determining several degrees of smoothing in the regression estimation. The insensitive zone e provides the required level of accuracy to approximate a function $f(x)$ by another function $f^*(x)$ such that the function $f(x)$ is situated in the e-tube of $f^*(x)$. The axis of the tube defines an e-approximation $f^*(x)$ of the function $f(x)$ (figure 2).

The choice of SVM as regressor is not only related to the robust character of its estimates, even when only sparse data is available, but also to the possibility of providing a particular interpretation to the support vectors. UMEL assumes that, in general, some of the support vectors of a regression represent the most difficult samples to regress and their coordinates in the input domain determine the exact location of special signatures in the signal.

Figure 3 shows a step function corrupted with pink noise and regressed with SVM. Some support vectors are located inside the smooth e-tube (ISV or internal support vectors) but others lay outside (ESV or external support vectors). Samples that become ESV are always on or outside the *e*-bounds and show special signatures, in this case a discontinuity.

Typically, in plasma physics, physical events are recognized by particular structural forms in the signals (patterns) such as spikes, transients or gradients. These patterns are signatures well located in the signals and can appear in waveforms or images. UMEL recognizes any kind of singular event through the external support vectors.

A new development with UMEL is the location of ELMs with temporal evolution signals of the JET database. ELMs are recognized by simultaneous abrupt changes in the temporal evolution of the Dα emission, the line integrated electron density through the plasma edge and the stored diamagnetic energy. UMEL determines the time instants of these singular footprints by means of the ESV (figure 4). A simultaneous occurrence of ESV in all signals is considered to be an indicator of the occurrence of an ELM.

It should be emphasized that exactly the same software is used to get the estimation regressions of any signals, irrespective of its magnitude, structural form and noise level. The only differences in the various versions of the algorithm are the regression parameters (regularization parameter, *e*-value and kernel parameter).

As an example of applications aimed at identifying time intervals of potential interest, the analysis of videos acquired by JET ultra high speed visible camera is presented in the following.
This camera can collect hundreds of kframes/s to analyse fast events. Some examples of the phenomena to be investigated with this diagnostic are plasma interactions with the in-vessel components, striations, filamentary structures and blobs in the plasma edge.

Each individual frame of a movie can be seen as a two-dimensional function where *x* and *y* are spatial coordinates and the amplitude of *f* at any point of coordinates $f(x, y)$ is the intensity or gray level of the image at that point (or pixel). Many of the aforementioned physical events produce peaked values of   over specific pixels. These peaks induce gradients in $f(x, y)$ and the gradients can be recognized by UMEL through the presence of ESV.

Variations in the number of ESV between consecutive frames denote the presence of images with different structural forms. This indicates that the plasma emission is changing and, therefore, this reflects how the plasma evolves in time.

Figure 5 shows the number of ESV versus the frame number in Pulse No: 69787 of JET. This has been calculated by applying UMEL image by image to the video-movie. The peaks with extreme number of ESV indicate singular points in the plasma evolution. They are exactly determined with a new application of UMEL to the waveform that represents the evolution of the number of ESV. The locations provided by UMEL are represented by red circles.

ACKNOWLEDGEMENTS

This work was partially funded by the Spanish Ministry of Science and Innovation under the Project No. ENE2008-02894/FTN. This work, supported by the European Communities under the contract of Association between EURATOM/CIEMAT, was carried out within the framework of the European Fusion Development Agreement. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

REFERENCES

[1]. “The Spider - A machine learning in Matlab". http://www.kyb.tuebingen.mpg.de/bs/people/spider/main.html. Max-Planck Institute for biological Cybernetics, Tuebingen, Germany.

[2]. V. Cherkassky, F. Mulier. “Learning from data“. John Wiley & Sons, Inc. (1998).

[3]. J. Vega, A. Murari, G. Vagliasindi, G.A. Rattá. “Automated estimation of L/H transition times at JET by combining Bayesian statistics and Support Vector Machines”. Nuclear Fusion (in press).

[4]. A.J. Meakins, D.C. McDonald. “The Application of Classification Methods in a Data Driven Investigation of the JET L-H Transition”. In submission to Plasma Physics and Controlled Fusion.

[5]. L.I. Kuncheva. “Combining pattern classifiers: Methods and Algorithms”. Wiley. (2004).

[6]. R.O. Duda, P. E. Hart, D.G. Stork. ”Pattern classification”. Second edition. Wiley-Interscience. (2001).

[7]. H.J. Zimmermann. “Fuzzy set theory and its applications”. Kluwer Academic Publishers, London (1991).

[8]. J. Ramírez, S. Dormido-Canto, J. Vega. “Automatic parallelization of classification Systems based on Support Vector Machines: comparison and application to the JET database”. (These proceedings).

[9]. J. Vega, A. Murari, S. González. “Universal method for automatic event location in waveforms and video-movies. Application to massive nuclear fusion databases”. In submission to Review of Scientific Instruments.

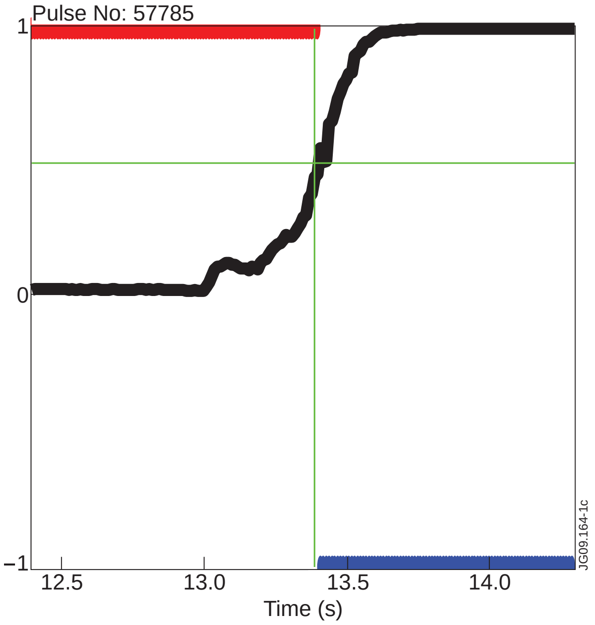

Figure 1: Automatic determination of the transition time with the combined classifier.

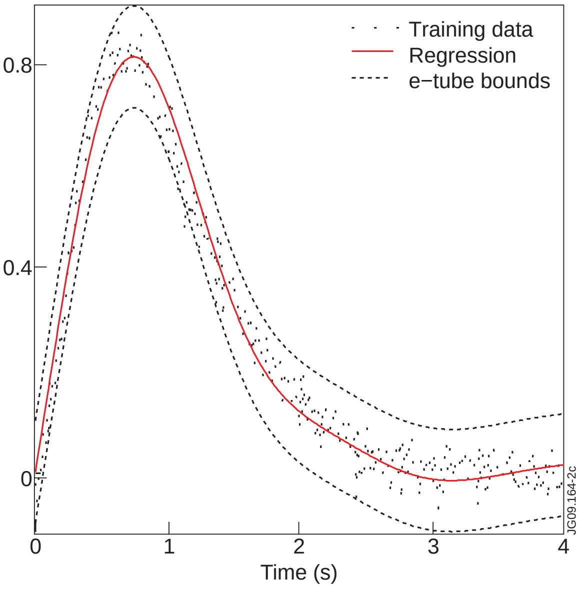

Figure 2: The solid line is the axis of the tube and defines the regression estimation.

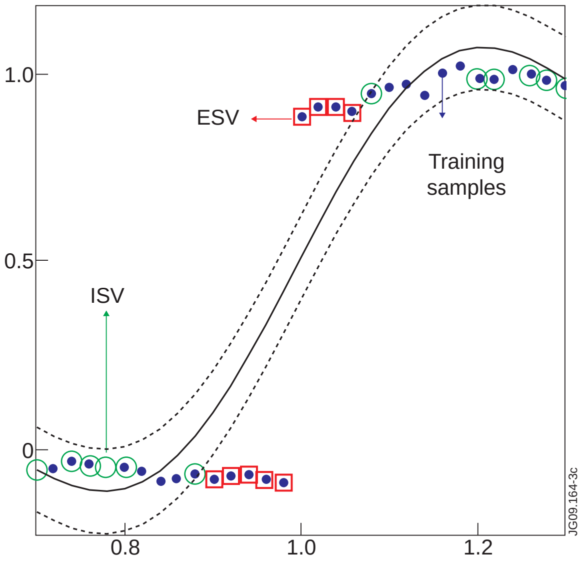

Figure 3: The dashed lines are the e-tube bounds and the plain line is the regression estimation.

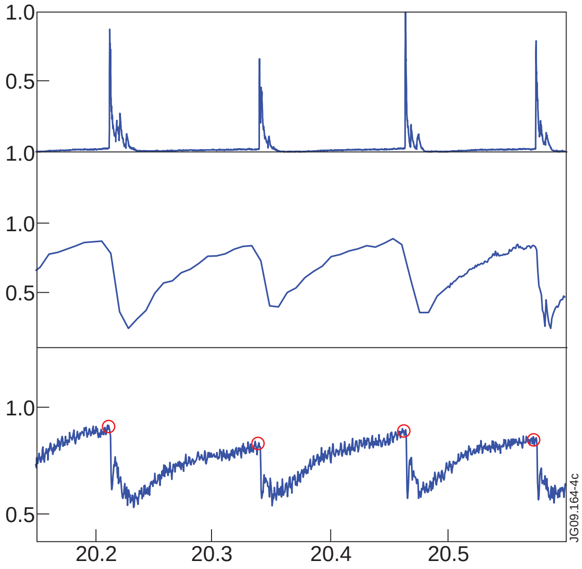

Figure 4: The ELM is located at the time instant when the diagmagnetic energy drops. Signals from top to bottom are: Dα, line integrated electron density and stored diamagnetic energy.

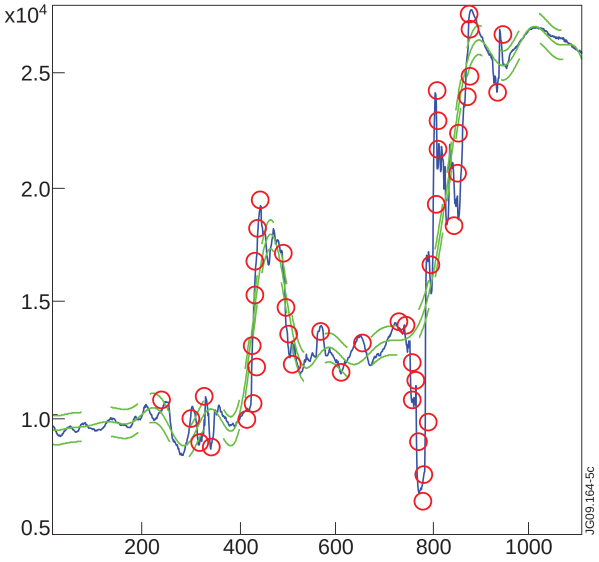

Figure 5: Time evolution of the ESV number for JET Pulse No: 69787.
The frames with an extreme variation in the number of ESV are shown in red.