Automated estimation of L/H transition times at JET by combining Bayesian statistics and support vector machines

IOP Publishing and International Atomic Energy Agency Nuclear Fusion

Nucl. Fusion **49** (2009) 085023 (11pp) doi:10.1088/0029-5515/49/8/085023

Automated estimation of L/H transition times at JET by combining Bayesian statistics and support vector machines

J. Vega[1], A. Murari[2], G. Vagliasindi[3], G.A. Rattá[1] and JET-EFDA Contributors[a]

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] Dipartimento di Ingegneria Elettrica Elettronica e dei Sistemi-Università degli Studi di Catania, 95125 Catania, Italy

Received 22 January 2009, accepted for publication 12 June 2009
Published 21 July 2009
Online at stacks.iop.org/NF/49/085023

Abstract
This paper describes a pattern recognition method for off-line estimation of both L/H and H/L transition times in JET. The technique is based on a combined classifier to identify the confinement regime (L or H) at any time instant during a discharge. The classifier is a combination of two different classification systems: a Bayesian classifier whose likelihood is computed by means of a non-parametric statistical classifier (Parzen window) and a support vector machine classifier. They are combined through a fuzzy aggregation operator, in particular the Einstein sum. The success rate achieved exceeds 99% for the L to H transition and 96% for the H to L transition. The estimation of transition times is accomplished by following the temporal evolution of the confinement regimes.

PACS numbers: 52.25.−b, 29.85.Fj, 07.05.Mh

(Some figures in this article are in colour only in the electronic version)

1. Introduction

Plasma evolution is not uniform during a tokamak discharge. Several and varied physical events appear as a consequence of the complex physics involved and the highly non-linear interactions among several physical processes. Various phenomena such as L/H transitions, edge localized modes (ELMs), internal transport barriers, disruptions or diverse types of instabilities can take place in different time instants in different discharges. Certain events can be induced in the pursuit of an experimental goal (for example L/H transitions and internal transport barriers) or, on the contrary, they can be present in an intermittent way as a result of the plasma natural evolution (for instance ELMs and magnetohydrodynamic instabilities). In any case, the creation of ad hoc databases to analyse the physics involved and determine the exact times when events occur is in general a difficult problem. These times are often estimated by relating multiple signals of a discharge in a manual way (typically by means of visual data analysis). Obviously, this procedure is not optimal at all, taking into account the large amount of data stored in present day fusion databases. Hence, the development of techniques for the automatic estimation of time instants associated with the occurrence of particular physical events is an urgent need to expedite the data analysis process and to guarantee unbiased results.

Particularly important is the requirement to create databases with enough statistical significance. For instance, studies related to disruptions, L/H transitions or ELMs require a big effort for the manual identification of the time instants where the events happen. Therefore, very often conclusions are drawn on the basis of a very limited number of discharges, whose analysis can be very accurate but the general relevance questionable. This paper proposes the use of pattern recognition techniques for the automated processing of big databases from a reduced number of examples. In particular, attention is focused on L to H and H to L transitions.

Pattern recognition techniques have recently been proposed for efficient data retrieval methods [1, 2]. Pattern-based data retrieval systems have been developed for the databases of the JET tokamak and the TJ-II stellarator [3]. In

[a] See the Appendix of F. Romanelli *et al* 2008 *Proc. 22nd IAEA Fusion Energy Conf. (Geneva, Switzerland, 2008).*

this paper, pattern recognition is used to classify the operational regime of JET at any time and to determine the time instants where transitions (both L/H and H/L) take place.

There exist some previous works devoted to developing classification systems to know the confinement mode [4–6]. The first one compares a couple of classification techniques (support vector machine (SVM) and fuzzy logic) for the JET database. The best rate of classification is 95.77%. The second develops a classification system to analyse possible operational modes in ITER. To this end, experimental signals from different tokamaks have been collected worldwide and are made available in the International Global H-Mode Confinement Database (http://efdasql.ipp.mpg.de/HmodePublic). The success rate of the classification process is 92.66%. In both cases, the analysis covers only the transition from L to H mode (and not the inverse H to L one) and exclusive regime identification aspects are treated. The third of the previous references describes a neural network classification technique to identify the nature of the dependence of the JET L–H and H–L boundary on local edge plasma parameters. The success rate achieved here is 98.86%.

In the above cases, the corresponding classification systems are trained using sets of scattered points at both sides of the transitions. This means that a sparse point population could be present just around the transitions and, therefore, big errors could arise if the preceding methods were used to determine the transition times. In contrast, this paper chooses feature vectors with a constant sampling period. This paper describes a method to determine the changes in confinement regime (not only from L to H but also from H to L) and, therefore, it can be applied to the entire temporal evolution of the discharges. To this end, several classification systems have been developed for both types of transitions. The ones obtaining the best success rates have then been used to determine the time instants in which the plasma transits from one confinement regime to another.

The L/H classification is a two class problem ($C_{\mathrm{L}}$and C_{H}). The classifier has to implement a decision algorithm whose result for an input feature vector must be $y = +1$ (which means L mode) or $y = -1$ (H mode). It is worth mentioning that traditional statistical classifiers based on parametric techniques are valid under the assumption that the forms of the underlying density functions are known (for instance, normal multivariate distributions with unknown mean and covariance matrix). However, the common parametric forms rarely fit the densities actually encountered in practice. In particular, all the classical parametric densities are unimodal (have a single local maximum), whereas many practical problems involve multimodal densities. Hence, in order to avoid unjustified assumptions about the form of the probability density functions, the approaches used in this paper for the L/H classification problem will be based on non-parametric statistical classifiers. Indeed, classification methods exist which do not require knowledge of the forms of underlying probability distributions, and in this sense they can be said to be non-parametric. SVM belongs to these types of classifiers. SVM is a universal constructive learning procedure based on statistical learning theory [7]. SVM relies on pre-processing the data to represent the feature vectors in a high dimensional space—typically of much higher dimensions than the original feature space. With an appropriate non-linear mapping to a sufficiently high dimensional space, data from two categories can always be separated by a hyper-plane (or decision function) [8].

An estimation of the probability of each classification can be derived from the decision function $D(\boldsymbol{u})$ provided by SVM. This can be done remembering that to classify a feature vector $\boldsymbol{u}$, only the sign of $D(\boldsymbol{u})$ is required. However, the absolute value of $D(\boldsymbol{u})$ provides the distance of the feature vector to the separating hyper-plane. This magnitude can be interpreted as a measure of certainty about its membership degree to the classes, in this paper C_{L} and C_{H}. The greater $|D(u)|$ the deeper is $\boldsymbol{u}$ in its corresponding confinement regime. Therefore, the probability of a classification can be identified through the decision function by applying the sigmoid function to $D(\boldsymbol{u})$. Due to the fact that the L mode is represented by $y = +1$, the probability of C_{L} can be expressed as

$$P_k(C_{\mathrm{L}}) = \frac{1}{1 + \exp[-kD(\boldsymbol{u})]}$$

and, therefore, $P_k(C_{\mathrm{H}}) = 1 - P_k(C_{\mathrm{L}})$. Different values can be assigned to the parameter k ($k_{\min} \leqslant k \leqslant k_{\max}$) and, hence, different probabilities are obtained.

A different classifier can be based on the Bayes decision theory. Given a classification task of M classes, ω_1, ω_2, …, ω_M, and an unknown pattern, which is represented by a feature vector $\boldsymbol{x}$, the Bayes rule states

$$P(\omega_k|\boldsymbol{x}) = \frac{\boldsymbol{p}(\boldsymbol{x}|\omega_k)\boldsymbol{P}(\omega_k)}{\boldsymbol{p}(\boldsymbol{x})}.$$

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

$$p(\boldsymbol{x}) = \sum_{i=1}^{M} \boldsymbol{p}(\boldsymbol{x}|\omega_i)\boldsymbol{P}(\omega_i).$$

It should be noted that $p(\boldsymbol{x})$ can be viewed merely as a scale factor that guarantees that the posterior probabilities sum to one. Therefore, the Bayes classification rule for the binary problem of assigning a feature vector $\boldsymbol{u}$ as L mode (C_{L}) or H mode (C_{H}) is reduced to

$$\begin{array}{ll} \boldsymbol{u} \in C_{\mathrm{L}} & \text{if } P\left(C_{\mathrm{L}}|\boldsymbol{u}\right) \geqslant P\left(C_{\mathrm{H}}|\boldsymbol{u}\right), \\ \boldsymbol{u} \in C_{\mathrm{H}} & \text{otherwise.} \end{array}$$

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 (see section 6). With regard to 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 0.5 for both cases.

The main goal of the paper is to combine the results of different classification methods to obtain better success rates [9, 10]. In particular, a combination of the SVM method with the Bayes formula is used. The former provides the probability that the plasma is in H mode (P_{SH}). It is computed

by means of the distance to the separating hyper-plane through the sigmoid function. The latter gives the posterior probability (Bayes rule) that the plasma is in H mode (P_{PH}) when the likelihood is computed with the Parzen window method and the prior probabilities are assumed to be 0.5 for each confinement regime. The fusion of both outputs is carried out by means of a fuzzy aggregation operator [11]. In particular, we have chosen the Einstein sum

$$S_{\mathrm{H}} = \frac{P_{\mathrm{SH}} + P_{\mathrm{PH}}}{1 + P_{\mathrm{SH}} P_{\mathrm{PH}}}.$$

The final decision of this hybrid classifier is based on the value of S_{H}:

$$\begin{array}{ll} \text{the plasma is in H mode} & \text{if } S_{\mathrm{H}} \geqslant 0.5, \\ \text{the plasma is in L mode} & \text{otherwise.} \end{array}$$

With regard to the structure of the paper, section 2 describes pattern recognition as a particular case of machine learning. Section 3 is devoted to providing the basic concepts and nomenclature that will be used in the remaining part of the paper. Section 4 explains, firstly, the JET datasets used for the transition time estimations. Secondly, it justifies the selection of the discriminating signals that best represent the L/H and H/L transitions. Thirdly, the normalization of the signals is explained. Sections 5 and 6 summarize different learning methods and their application to the recognition of L and H confinement regimes in JET. Section 7 is the most important section of this paper. It shows how to combine the Bayes decision rule (the likelihood is computed according to the Parzen window estimator) and SVM classifiers through a fuzzy aggregation operator (specifically the Einstein sum) to obtain better success rates in the classification of L/H modes. Section 8 reports the transition time determination from classification systems. Finally, section 9 gives a short discussion of the method and suggests future work.

All computations have been performed with Matlab[4] and, in particular, the SVM implementation has been 'The Spider' software included in public licensed environments for Matlab [12].

2. Mathematical description of learning systems

Machine learning can be considered as the process whereby the dependence between quantities is determined using a limited (sometimes very limited) number of observations. From a mathematical point of view, the general model of learning from data can be described through three components [7]: first, a generator (G) of vectors $\boldsymbol{x} \in \mathbb{R}^n$, drawn independently from a stationary but unknown pdf $F(\boldsymbol{x})$; second, a supervisor (S) who returns an output value y to every input vector $\boldsymbol{x}$, according to a conditional distribution function $F(y|\boldsymbol{x})$, also given but unknown; third, a learning machine (LM) capable of implementing a set of functions $f(\boldsymbol{x}, \alpha)$, $\alpha \in \Lambda$, where Λ is a set of parameters.

The problem of learning is that of choosing from the given set of functions $f(\boldsymbol{x}, \alpha)$, $\alpha \in \Lambda$, the one that best approximates the supervisor's response. The selection of the desired function is based on a training set of ℓ independent and identically distributed observations $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_\ell, y_\ell)$ drawn according to $F(\boldsymbol{x}, y) = F(\boldsymbol{x})F(y|\boldsymbol{x})$.

[4] http://www.mathworks.com/.

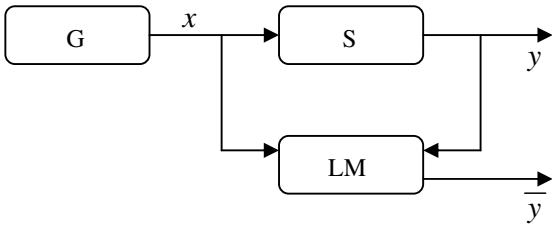

Figure 1. A sample generator (G) sends samples (x) to a supervisor system (S) that sets a value (y) for each sample. The pairs (x, y) are introduced to a learning machine (LM) with the aim of learning. The goal is that LM can return values ($\overline{y}$) for each input x that are close to the supervisor's response.

Figure 1 represents a model of learning from examples. During the learning process, the learning machine observes the pairs ($\boldsymbol{x}$, y) (the training set). After training, the machine must return a value $\overline{y}$ for any given $\boldsymbol{x}$. The goal is to return a value $\overline{y}$ that is as close as possible to the supervisor's response y.

This formulation of the learning problem is rather broad. It encompasses many specific tasks such as pattern recognition (or classification, i.e. estimation of class decision boundaries), regression (or estimation of an unknown continuous function from noisy samples) and probability density estimation. In contrast to the classical statistics developed for large samples, the theory of learning systems was developed for small data samples. This characteristic implies that learning theory is optimal for physical systems that have to be described with an insufficient number of data. This is a very important reason to use the learning theory in plasma physics because a typical constraint in fusion experimental environments is the restricted amount of information.

The creation of databases to provide L/H transition times is treated here as a pattern recognition problem. The objective is to minimize the probability of classification error (L mode or H mode) when the probability measure $F(\boldsymbol{x}, y)$ is unknown, but the data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_\ell, y_\ell)$ are given.

3. Introduction to classification methods

Generally speaking, a classification system can be seen as a set of coordinated methods for object description and classification. Object description is a way of representing objects to be managed by computers. This description must include as much *a priori* knowledge as possible about the problem to solve. Only a limited number of features or attributes (say n) of the objects are of distinctive nature for the problem under consideration. The numerical values of the n features pertaining to each object are represented as a vector $\boldsymbol{x} \in \mathbb{R}^n$ called the feature vector. To achieve a suitable representation, objects are usually pre-processed in an iterative way to find an adequate description for the problem to solve.

With regard to JET transitions, feature vectors are the inputs to a supervisor system (figure 1) that establishes a label for the feature vector. In this case, the supervisors are specialists who have determined the transition times of a reduced number of discharges. The labels used in this work

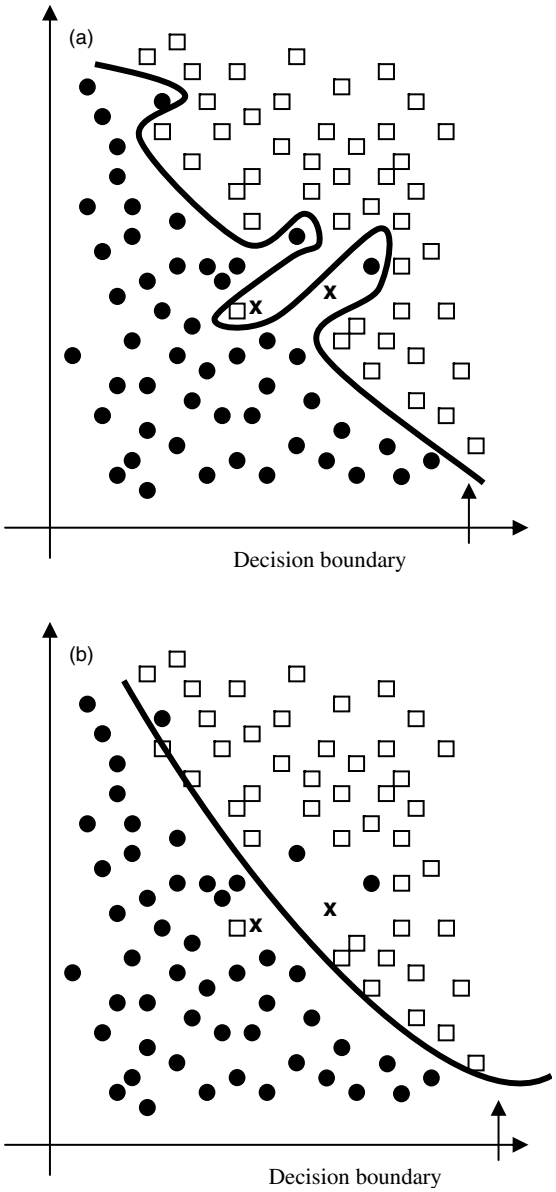

Figure 2. Figure 2(*a*) shows overfitting. The generalization capability is not optimal. Where are classified the points marked as *x*? The classifier in figure 2(*b*) presents a better generalization property.

to distinguish the confinement regimes are $y = +1$ for the L mode and $y = -1$ for the H mode.

The feature vectors that have been labelled by the supervisor form the training set. They are then used as inputs to the learning machine with the aim of generating a classifier (model) that is able to classify new objects without the help of the supervisor. The process of model generation is called 'training'. At this point, it is important to emphasize that in our context learning does not mean 'learning by heart' (in fact, any computer can memorize). The challenge is to classify new objects (represented by their corresponding feature vectors) that can be different from the ones used for the training. The system capability of classifying new objects that were not used in the training phase is called generalization (figure 2).

After building the model in the training process, it is necessary to quantify the generalization power of the classifier. An obvious criterion could be the training error, i.e. the average loss incurred in classifying the set of training samples. It should be emphasized right away, however, that despite the attractiveness of this criterion, it is fraught with problems. Since the goal is to classify novel test patterns, a small training error does not guarantee a small test error. For this reason, a test set is used to determine an independent error rate. The test set is made of feature vectors that have been labelled by the supervisor but are completely new to the model. The success rate of the learning machine in classifying this set of new objects determines the goodness of the model. Typically, poor results make it necessary to modify some or all of the previous steps: feature selection, number of training data and classifier model.

4. Datasets for the L/H transitions

The goal of this work is the off-line identification of the transition times. To this end, the coordinates of each feature vector will be synchronous samples of different time series data. The temporal segment considered for the training process has been 2 s long, i.e. the time instants of the first and last feature vectors in each discharge correspond to a time of 1 s before and after the transition, respectively. The time interval between feature vectors is 10 ms.

The starting point of this research to determine the transition from the low (L) to the high (H) regime has been a set of 50 discharges between shots 52211 and 62723 (with the septum in the JET divertor configuration) whose transition times have been established by experts with a high degree of confidence [6]. Only for 42 shots all signals are available during the 2 s segments around the transition to create the feature vectors of the learning system. 80% of them are used in a random way for training purposes (33) and the rest constitute the corresponding test set. The test set allows estimation of the success rate of the classification system in deciding the regime associated with each feature vector.

Regarding the H/L transition, 48 shots have been considered in the previous range of discharges (also with the septum) and again, the transition times have been determined by specialists [6]. In contrast to the L/H case, the time instant determination is more difficult. Only 38 discharges have all the samples during 2 s around the estimated transition instant and 80% of them (30 shots chosen at random) are used for training and the remaining ones for test.

The training and test datasets cover a wide range of discharges. For example, the toroidal magnetic field is between 1.8 and 3.4 T, the plasma current between 1.2 and 4.3 MA and the edge line integrated density between 2.2×10^{19} and $1.1 \times 10^{20}\,\mathrm{m}^{-2}$.

For the JET L/H transitions, 35 signals have been selected as candidates to provide discriminant characteristics for the pattern recognition problem. They also include geometrical parameters to take into account the position/shape of the plasma inside the vacuum vessel. The signals have been previously provided as inputs to a tree structured methodology in classification: CART (Classification and Regression Trees) [13]. The CART outputs provide as result the variable ranking

Table 1. List of most important signals to describe the confinement regimes with the Septum in the JET divertor configuration.

Transition	Signal	Signal description
L/H	Bndiam	Beta normalized with respect to the diamagnetic energy
	Ptot	Total heating power
	Wmhd	Magnetohydrodynamic energy
	RXPL	R coordinate lower XP
	ZXPL	Z coordinate lower XP
	Bt80	Axial toroidal magnetic field at a psi = 0.8 surface
H/L	Bndiam	Beta normalized with respect to the diamagnetic energy
	Bt	Toroidal magnetic field
	FDWDT	Time derivative of diamagnetic energy
	Q95	Safety factor
	RXPL	R coordinate lower XP
	ZXPL	Z coordinate lower XP

of the most relevant signals for the classification problem. For JET discharges with the septum in the divertor configuration, six signals are determined as the most important ones for the L/H transition (table 1) and a different set of six variables for the H/L transition.

Therefore, feature vectors to represent the confinement regime belong to a six-dimensional space, where each coordinate is the value of one of the physical quantities that appear in table 1. It should be highlighted that these quantities correspond to the same time instant. Hence, the process of generating the feature vectors implies a data pre-processing step consisting of interpolating the time series data to a common set of time instants. This step is required to overcome the difficulty that the signals in JET are sampled at different frequencies.

Finally, an important issue has to be taken into account. All the coordinates of the feature vectors must weight in the same way for the model generation. Hence, to avoid coordinates represented by big numbers having an excessive influence due to their absolute values and not to their information content, a suitable normalization process is mandatory. The normalization process is carried out for each physical quantity in the training process. Given a quantity M, its corresponding coordinate in the feature vector is

$$\frac{M - \min(M)}{\max(M) - \min(M)},$$

where $\min(M)$ and $\max(M)$ are, respectively, the minimum and maximum values of M for all the discharges considered in the training set. This transformation generates coordinates between 0 and 1 thereby ensuring the same relative importance for all quantities.

5. Support vector machines

Given a training set of ℓ samples $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_\ell, y_\ell)$, $\boldsymbol{x}_i \in \mathbb{R}^n$, for a binary classification problem (i.e. $y_i \in \{+1, -1\}$), SVM estimates the following decision function:

$$D(\boldsymbol{x}) = \sum_{i=1}^{\ell} \alpha_i y_i H(\boldsymbol{x}_i, \boldsymbol{x})$$

where $H(\boldsymbol{x}_i, \boldsymbol{x})$ is a kernel function [14] and the parameters $\alpha_i, i = 1, \ldots, \ell$ are the solutions of the following quadratic optimization with linear restrictions:

maximization of the functional

$$Q(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j H(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

subject to the constraints

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0, \qquad 0 \leqslant \alpha_i \leqslant \frac{C}{\ell}, \qquad i = 1, \ldots, \ell,$$

where C is a regularization parameter [14].

The data points $\boldsymbol{x}_i$ associated with the nonzero α_i are called support vectors. Once the support vectors have been determined, the SVM decision function has the form

$$D(\boldsymbol{x}) = \sum_{\text{support vectors}} \alpha_i y_i H(\boldsymbol{x}_i, \boldsymbol{x}),$$

where $D(\boldsymbol{x})$ is the distance from $\boldsymbol{x}$ to the hyper-plane that separates the two classes and, hence, the hyper-plane points satisfy $D(\boldsymbol{x}) = 0$. It should be noted that $D(\boldsymbol{x})$ actually is only proportional to the real distance. However this is not a problem, because all data points are rescaled by the same normalization factor.

The rule to clarify a feature vector $\boldsymbol{u}$ as L mode (class C_{L}) or H mode (class C_{H}) is given by

$$\begin{array}{ll} \boldsymbol{u} \in C_{\mathrm{L}} & \text{if } sgn(D(\boldsymbol{u})) \geqslant 0, \\ \boldsymbol{u} \in C_{\mathrm{H}} & \text{otherwise,} \end{array}$$

where sgn(t) is the sign function, i.e.

$$sgn(t) = \begin{cases} 1 & \text{if } t \geqslant 0 \\ -1 & \text{if } t < 0 \end{cases}.$$

5.1. L/H transitions

As mentioned previously, 33 discharges out of 42 have been chosen in a random way to form the training set in order to generate the model of the L/H transition. Each discharge contributes with 200 feature vectors (50% of each confinement regime and 10 ms of the sampling period between vectors) in the time segment (in s) $(-1, 1)$, where times are related to the transition times estimated by specialists. This set-up implies 6600 feature vectors for the training phase.

The test set is made up of 1800 feature vectors and 246 different classification systems have been developed. The performance of each model is evaluated according to the success rate achieved with the test set.

The first classification system uses a linear kernel $(H(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x} \cdot \boldsymbol{x}')$ and obtains a success rate of 94.91% with the training set. The test set yields a success rate of 96.61% (see table 2).

The SVM model with the linear kernel also gives very important extra information regarding the L/H transitions. The equation of the separating hyper-plane (normalized data) is

$$-7.3^{*}\text{Bndiam} + 17.8^{*}\text{Ptot} - 54.8^{*}\text{Wmhd} + 1.5^{*}\text{RXPL} + 0.7^{*}\text{ZXPL} - 10.3^{*}\text{Bt80} + 13.3 = 0.$$

Table 2. Success rates achieved with the test set and different classifiers: SL (SVM, linear kernel), SR (245 classifiers based on SVM and RBF kernels), PW (60 models based on Parzen window estimators) and BSR (Bayes/SVM (RBF kernel) combination).

	L/H transitions	H/L transitions
SL	96.61%	89.38%
SR (245)	99.11%	94.44%
	$\sigma = 10$	$\sigma = 39$
PW (60)	98.61%	95.88%
	$0.5h_{j,\mathrm{REF}} \leqslant h \leqslant 0.71h_{j,\mathrm{REF}}$	$h = 0.95h_{\mathrm{REF}}$
BSR	99.22%	96.31%
	$\sigma = 10$	$\sigma = 39$
	$h = 1.55h_{\mathrm{REF}}$	$h = 0.79h_{\mathrm{REF}}$
	$k = 2$	$k = 100$

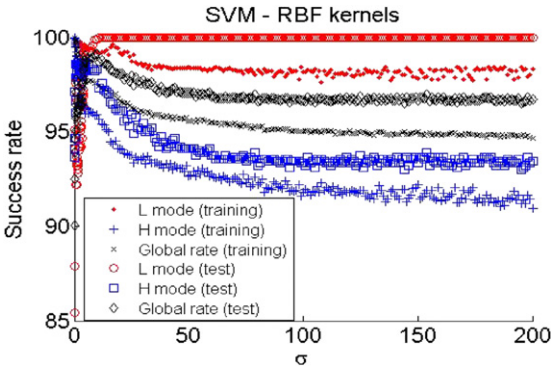

Figure 3. Success rates (%) in L/H transitions for the training and test datasets with SVM and a RBF kernel.

This equation relates the physical magnitudes among them in such a way that it could be possible to establish threshold values for the L/H transition with individual variables, assuming that the rest of them are known.

On the other hand, due to the fact that the distance to the separating hyper-plane is expressed as

$$D(\boldsymbol{x}) = -7.3^{*}\mathrm{Bndiam} + 17.8^{*}\mathrm{Ptot} - 54.8^{*}\mathrm{Wmhd} + 1.5^{*}\mathrm{RXPL} + 0.7^{*}\mathrm{ZXPL} - 10.3^{*}\mathrm{Bt80} + 13.3$$

it implies that given the experimental values of these variables, JET can be operated in deep either L or H mode (it should be recalled that according to the sign criteria of this paper, a large positive value of $D(\boldsymbol{x})$ means a deep L confinement regime whereas a large negative quantity denotes a deep H confinement mode).

As mentioned previously, 245 additional classification systems have been developed. They have been based on a radial basis function (RBF) kernel:

$$H(\boldsymbol{x}, \boldsymbol{x}') = \exp\left\{-\frac{|\boldsymbol{x} - \boldsymbol{x}'|^2}{\sigma^2}\right\}.$$

This kernel requires the *a priori* selection of its scale factor σ. The scale factor defines the size (or width) of the region around $\boldsymbol{x}$ for which H is large. The range of σ was $0.1 \leqslant \sigma \leqslant 200$ and the success rates are shown in figure 3. With the training set, a 100% success rate is achieved for $\sigma = 0.1$ and a value of 99.11% is obtained with the test set for $\sigma = 10$ (table 2).

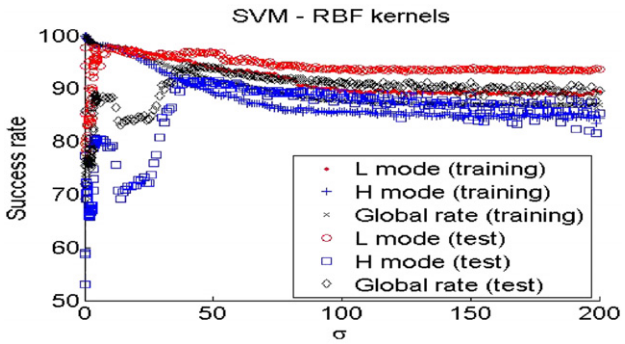

Figure 4. Success rates (%) in H/L transitions for the training and test datasets with SVM and a RBF kernel.

5.2. H/L transitions

Here, the training set consists of 6000 feature vectors from 30 discharges chosen randomly from a set of 38 shots and under the same conditions established for the L/H case (200 feature vectors per discharge, 50% of each confinement mode, 2 s length segments and 10 ms of sampling period). Under these conditions, results are shown in figure 4 and table 2. With a linear kernel, success rates are 86.63% and 89.38% for the training and test sets, respectively.

Now, the hyper-plane equation is

$$-24.8^{*}\mathrm{Bndiam} + 0.7^{*}\mathrm{Bt} - 7.8^{*}\mathrm{FDWDT} - 1.8^{*}\mathrm{Q95} - 5.8^{*}\mathrm{RXPL} + 4.5^{*}\mathrm{ZXPL} + 16.6 = 0.$$

Of course, the same considerations performed with the L/H case on both variable relations and confinement depths are applicable now.

With RBF kernels, the maximum success rate is again 100% with $\sigma = 0.1$ for the training set and 94.44% ($\sigma = 39$) for the test set.

6. Non-parametric statistical classifier: Parzen window

Among the non-parametric probability density estimators, the Parzen window method is the most popular. It is described in great detail in [8].

For univariate distributions, the kernel estimator is given by

$$p(x) = \frac{1}{\ell h}\sum_{i=1}^{\ell} K\left(\frac{x - X_i}{h}\right),$$

where ℓ is the number of samples, h the window width and the function $K(t)$ is called a kernel. The above equation expresses the estimate of $p(x)$ as an average of functions of x and the samples X_i. In essence, $K(t)$ is used for interpolation and each sample contributes to the estimate in accordance with its distance from x. $K(t)$ is itself a density function which satisfies $K(t) \geqslant 0$ and $\int K(t)\mathrm{d}t = 1$. Examples of kernels for density estimation appear in table 3.

The parameter h determines the amount of smoothing in the estimate $p(x)$. A small value of h yields a rough curve, while a large value of h yields a smoother curve. Detailed analyses about the parameter h are out of the scope of this paper, but it is necessary to mention its importance because

Table 3. Examples of kernels for density estimation.

Kernel name	Equation
Triangle	$K(t) = (1 - \|t\|) - 1 \leqslant t \leqslant 1$
Epanechnikov	$K(t) = \frac{3}{4}(1 - t^2) - 1 \leqslant t \leqslant 1$
Biweight	$K(t) = \frac{15}{16}(1 - t^2)^2 - 1 \leqslant t \leqslant 1$
Triweight	$K(t) = \frac{35}{32}(1 - t^2)^3 - 1 \leqslant t \leqslant 1$
Normal	$K(t) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-t^2}{2}\right) - \infty \leqslant t \leqslant \infty$

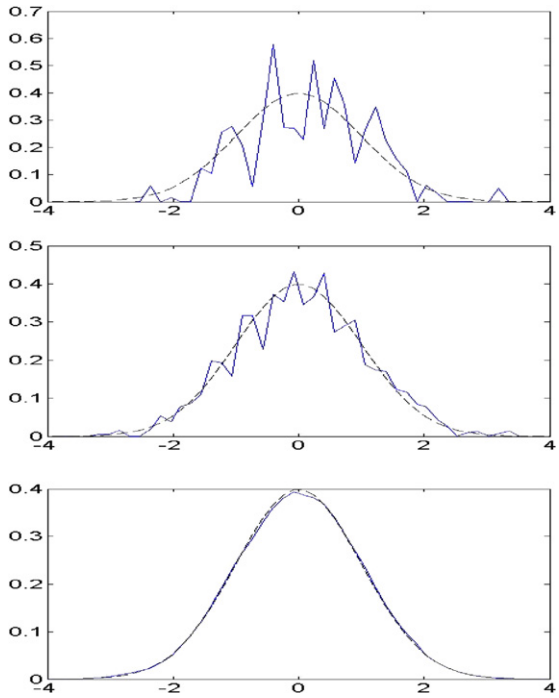

Figure 5. The top figure computes the estimator with a small window width ($h = 10^{-4}$). This value is increased in the central case ($h = 10^{-3}$) and finally, in the bottom figure, it reaches an optimal width given by the normal reference rule ($h = 0.121$, where σ was estimated as the standard deviation of the samples).

different values of h yield different classifiers. In fact, it is known that the choice of the smoothing parameter h is more important than the choice of the kernel. For a normal density kernel, the width h can be selected using the following criterion (called the 'normal reference rule') [15]:

$$h_{\mathrm{REF}} \approx 1.06\sigma\,\ell^{-1/5},$$

where a suitable estimate for σ is the standard deviation.

Figure 5 gives an example of the influence of h on the estimation of a normal distribution $N(0, 1) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2}\right]$, $-\infty < x < \infty$. The dashed line is the $N(0, 1)$ distribution to be estimated from a set of random samples (derived from the same distribution) and the continuous line is the Parzen window estimator with a Gaussian kernel.

In the cases where each observation is a n-dimensional vector, $\boldsymbol{x} \in \mathbb{R}^n$, the simplest case for the multivariate kernel estimator is the product kernel. The product kernel is

$$p(\boldsymbol{x}) = \frac{1}{\ell h_1 \ldots h_n} \sum_{i=1}^{\ell} \left\{ \prod_{j=1}^{n} K\left(\frac{x_j - X_{ij}}{h_j}\right) \right\},$$

where X_{ij} is the jth component of the ith sample [15]. Note that this is the product of the same univariate kernel, with a (possibly) different window width in each dimension. Since the product kernel estimate comprises univariate kernels, it is possible to use any of the kernels that have been mentioned previously. If the normal kernel is used, a normal reference rule for the multivariate case is [15]

$$h_j = \left(\frac{4}{\ell\,(n+2)}\right)^{\frac{1}{n+4}} \sigma_j, \qquad j = 1, \ldots, n \qquad (1)$$

and again, a suitable estimate for σ_j can be used. If there is any skewness or kurtosis evident in the data, then the window widths should be narrower.

By choosing Gaussian kernels (that typically provide very good results) for the L/H transitions, the probability density function can therefore be expressed as

$$p(\boldsymbol{x}) = \frac{1}{\ell(2\pi)^{n/2} \prod_{k=1}^{n} h_k} \sum_{i=1}^{\ell} \exp\left\{ -\frac{1}{2} \sum_{j=1}^{n} \left(\frac{x_j - X_{ij}}{h_j}\right)^2 \right\}.$$

Once the probability density function has been estimated for both the L regime ($p_{\mathrm{L}}(\boldsymbol{x})$) and the H regime ($p_{\mathrm{H}}(\boldsymbol{x})$), the classification of a feature vector $\boldsymbol{u}$ as L mode (class C_{L}) or H mode (class C_{H}) is carried out by means of the following decision rule:

$$\begin{array}{ll} \boldsymbol{u} \in C_{\mathrm{L}} & \text{if } p_{\mathrm{L}}(\boldsymbol{u}) \geqslant p_{\mathrm{H}}(\boldsymbol{u}), \\ \boldsymbol{u} \in C_{\mathrm{H}} & \text{otherwise.} \end{array}$$

A total number of 60 classifiers (models) have been generated for each training set. They differ in the window width parameter. In order to cover a wide collection of values, the range of h explored for each component is

$$0.5h_{j,\mathrm{REF}} \leqslant h \leqslant 2.05h_{j,\mathrm{REF}},$$

where $h_{j,\mathrm{REF}}$ is given by the normal reference rule and it is computed according to equation (1). The estimate for σ is the standard deviation of the training data in each one of the 6 feature vector components.

It should be noted that the classification results do not change at all by considering the Bayes decision rule. The only difference is that the posterior probabilities are obtained to be used later with the hybrid classifier. These posterior probabilities are computed from

$$P\left(C_j|\boldsymbol{x}\right) = \frac{p\left(\boldsymbol{x}|C_j\right) 0.5}{p\left(\boldsymbol{x}\right)}, \qquad j = \{\mathrm{L}, \mathrm{H}\},$$

where

$$p(\boldsymbol{x}) = 0.5\left[p\left(\boldsymbol{x}|C_{\mathrm{L}}\right) + p\left(\boldsymbol{x}|C_{\mathrm{H}}\right)\right].$$

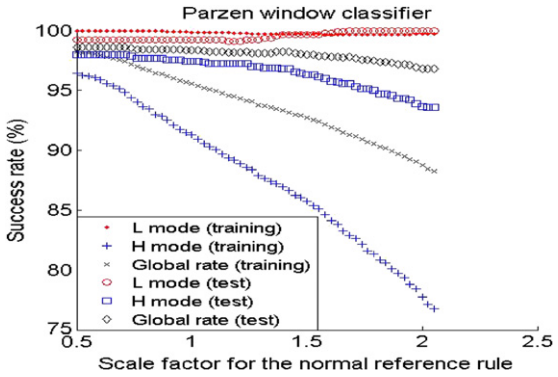

Figure 6. Parzen window classifier: success rate with 60 models for the L/H transition.

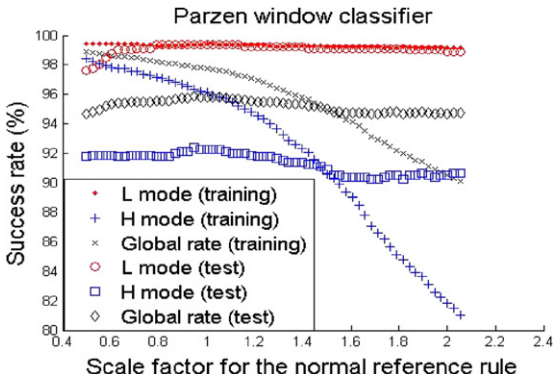

Figure 7. Parzen window classifier: success rate with 60 models for the H/L transition.

6.1. L/H transitions

By choosing the same training/test sets under identical conditions than in the case of the SVM, classifiers based on the Parzen window method have been developed. Figure 6 shows the success rates obtained with the 60 classifiers. The maximum success rate with the training set is 98.21% and it appears with the first h value ($h = 0.5h_{j,\mathrm{REF}}$). The test set achieves a success rate of 98.61% for the interval $0.5h_{j,\mathrm{REF}} \leqslant h \leqslant 0.71h_{j,\mathrm{REF}}$ (see table 2).

6.2. H/L transitions

With the same training/test datasets described in the SVM section for the H/L transition, the success rates are again reported in table 2. Figure 7 shows the performances of the 60 classification systems with both the training set and the test set. The former gives a success rate of 98.92% with $h = 0.5h_{j,\mathrm{REF}}$ and the latter 95.88% with $h = 0.95h_{j,\mathrm{REF}}$.

7. Combination of Bayesian and SVM

The combined classifier (SVM + Bayes/Parzen) by means of a fuzzy aggregation operator has been explained in the introduction section. For the SVM classifier, a set of 199 k values have been considered (in the range between $k_{\min} = 0.01$ and $k_{\max} = 100$) to assign a probability to the classifier through the sigmoid function. The selection of the parameter k is performed on the basis of the best success rate achieved with a test set as explained below.

The process of building the hybrid classifier is summarized in the following pseudo code:

Select a training set.
Select a test set.
Develop different SVM classifiers: several kernels and diverse kernel parameters.
Select the SVM model which provides the best success rate with the test set.
Develop different Parzen windows classifiers with several values of the parameter h ($h_{\min} \leqslant h \leqslant h_{\max}$).
Determine h and k.

LOOP (Parzen window) $h_{\min} \leqslant h \leqslant h_{\max}$
compute for the test set $p_h\,(\boldsymbol{u}|C_{\mathrm{L}})$ and $p_h\,(\boldsymbol{u}|C_{\mathrm{H}})$.
compute the posterior probabilities $P_h\,(C_{\mathrm{H}}|\boldsymbol{u})$ for the test set
LOOP (k parameter) $k_{\min} \leqslant k \leqslant k_{\max}$
compute for the test set $P_k\,(C_{\mathrm{H}})$
compute for the test set

$$S_{\mathrm{H}} = \frac{P_k\,(C_{\mathrm{H}}) + P_h\,(C_{\mathrm{H}}|\boldsymbol{u})}{1 + P_k\,(C_{\mathrm{H}})\,P_h\,(C_{\mathrm{H}}|\boldsymbol{u})}$$

classify according to $\boldsymbol{u} \in C_{\mathrm{H}}\,(C_{\mathrm{L}})\ \ iif$ $S_{\mathrm{H}} \geqslant (<)0.5$
compute success rate $S(h, k)$ for the test set
END LOOP (k parameter)
END LOOP (Parzen window)

Determine h^* and k^* that correspond respectively to the h and k parameters which maximize $S(h, k)$.

So, the combined classifier is determined by three elements: the SVM model providing the best success rate with the test set together with the h^* and the k^* parameters found in the above nested loops.

7.1. L/H transitions

Focusing the attention on the test set, table 2 shows the success rate (99.22%) achieved with a classification system that results from the combination of the Parzen window method using the Bayes Theorem and SVM (RBF kernel).

7.2. H/L transitions

Table 2 also shows the results for H–L transitions with the hybrid method. It allows achieving 96.31% as success rates for the combination of the Parzen estimator with Bayes rule and SVM (RFB kernel).

8. Transition times determination

According to the results shown in the above sections, the best classifiers for L/H and H/L transitions have been obtained using the Einstein sum to combine the SVM estimates with the probability calculated through the Parzen window method and the Bayes formula. Therefore, this approach will be used to determine the transition times.

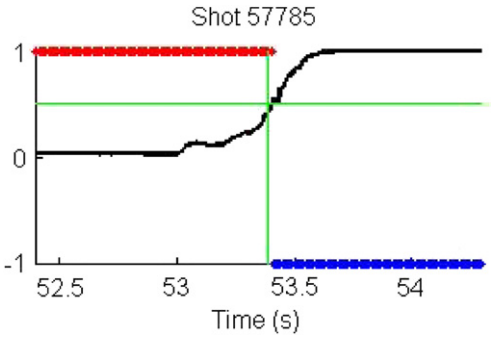

Figure 8. Transition time estimation with the combined classifier.

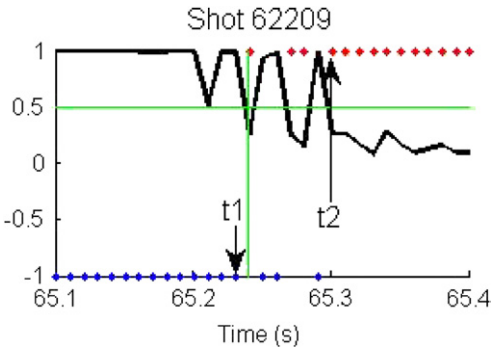

Figure 9. Example of transition that is not defined within a single point.

Given a shot number, the feature vectors between two arbitrary time instants are generated. It should be noted that the method does not require a constant sampling time to determine the transition instant. However, in the results shown here, again a 10 ms period was selected. The feature vectors are sent in time ascending order to the classifier.

Figure 8 shows an example of a time estimate that corresponds to a transition from the L to H regime. The vertical line (at time instant 53.39 s) represents the transition time determined by 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 (H) mode, respectively. The transition is estimated at 53.406 ± 0.005 s. The difference between the time determined by experts and the time deduced by the combined classifier is (in absolute value) 0.016 ms.

Figure 9 shows a rare evolution in which the H/L transition does not happen only once in the selected interval. The transition time is estimated as the average value between t_1 and t_2, i.e. the average value between the last H mode time slice before the first L mode time slice and the first L mode time slice after the last H mode time slice. In this example the estimated time is 65.266 ± 0.026 s. This has been the criterion for the estimations of L/H and H/L transitions times when uncertainties appear around the estimation. Further investigation is required to determine whether these rare cases are multiple transitions between the L and H modes or whether they are spurious oscillations of the classifier output due to the uncertainties in the measurements.

Table 4. Estimation of L/H transition times: TTE: transition time estimated by experts (unknown error bars), TH: Transition time estimated by the hybrid classifier and the last column is the difference (in absolute value) between TTE and TH.

Shot	TTE (s)	TH (s)	|TTE–TH| (s)
58259	58.57	58.73 ± 0.01	0.16
58404	53.23	53.29 ± 0.01	0.06
58760	59.43	59.28 ± 0.16	0.15
60592	50.18	50.06 ± 0.06	0.12
60682	53.26	53.09 ± 0.06	0.17
62549	52.70	52.75 ± 0.01	0.05
62557	52.54	52.56 ± 0.01	0.02
62558	52.70	52.69 ± 0.01	0.01

Table 4 sums up the L/H transition time estimations of 8 discharges. These shots correspond to a new dataset of L/H transitions that excludes the discharges used in the training and test phases. They are removed to avoid any bias in the presentation of results. The mean value of uncertainty in the estimation of the transition time is 40 ms, which implies an average error of 4 sampling periods. On the other hand, the mean value of the differences between the expert estimations and the deduced times (in absolute value) is 90 ms.

Regarding the H/L transition, with a new dataset of 10 discharges, the uncertainty average value around the transition is 170 ms (8 samples and a half at each side of the estimated time) whereas the mean value of the differences is 250 ms.

9. Discussion and future work

This paper shows that the combination of classifiers allows achieving higher success rates in the L/H pattern recognition problem. In particular, a specific mixture of SVM and the Bayes rule (with the likelihood computed with the Parzen window estimator) has been used. It provides the best performance in both the classification of the confinement regime and the estimate of the transition time. Table 2 summarizes the success rates with the test sets that confirm the increased success rates with the proposed approach.

The L to H case gets better results (99.22% of success rate) than the H to L (96.31% of success rate). This is a direct consequence of the uncertainties in the determination of transition times by specialists.

At this point, it is important to state whether the slightly greater success rates are significant improvements. To this end, let us consider sets of symmetric temporal intervals around the transitions: $[-1, -t_{\text{edge}}] \cup [t_{\text{edge}}, +1]$ (times are in s and are related to the transition time). A particular interval is the one having $t_{\text{edge}} = 0$, which obviously is the segment $[-1, 1]$. Table 5 gives the success rates for L to H transitions with individual classifiers. Also, it illustrates that except in time instants very close to the transition (less than 100 ms), the success rate is 100%. This means that the classifiers identify with the maximum probability the regime mode, i.e. the plasma confinement regime is known with probability 1. However, in times closer to the transition, misclassifications can occur. Table 6 shows the equivalent information for H to L transitions.

Table 5. Success rates of individual classifiers (SL, SVM with linear kernel; SR, SVM with RBF kernel; B, Bayes decision rule with Parzen window) for L to H transitions and different symmetric temporal intervals around the transition.

SL (%)	SR (%)	B (%)	t_{edge} (ms)
100	100	100	600
100	100	100	500
100	100	100	400
100	100	100	300
100	100	100	250
99.86	100	100	200
99.61	100	100	150
98.95	100	100	100
98.39	99.94	100	75
97.78	99.94	99.82	50
96.61	99.11	98.61	0

Table 6. Success rates of individual classifiers (SL, SVM with linear kernel; SR, SVM with RBF kernel; B, Bayes decision rule with Parzen window) for H to L transitions and different symmetric temporal intervals around the transition.

SL (%)	SR (%)	B (%)	t_{edge} (ms)
97.38	100	100	500
96.56	99.79	100	400
97.95	99.91	100	300
97.25	99.83	100	250
96.17	98.67	100	200
94.34	97.43	100	150
93.96	97.09	99.72	125
92.99	96.46	99.31	100
91.78	95.53	97.63	50
89.38	94.44	95.88	0

In this case, the results are reproduced but with a greater value of t_{edge}.

Therefore, in both cases, feature vectors can be misclassified as they approach the transition. Taking this fact into account and bearing in mind that initial error bars on the times estimated by the expert are unknown, an interpretation about the improvement in the time estimations with the hybrid method can be provided.

The training and test datasets used in this work are made up of feature vectors in the temporal segment with $t_{edge} = 0$. In this interval, 200 feature vectors per discharge are used. First of all, it should be noted that the success rates are really very high in both individual classifiers and hybrid classifiers. Second, and according to tables 5 and 6, the misclassifications always take place around the transition points. Third, focusing the attention on the L/H case, a success rate of 99.11% with SVM means that, on average, 2 points per discharge are misclassified. With the Bayes/Parzen classifier, the rate is 98.61%, which implies that 3 points are misclassified. The combined method achieves a success rate of 99.22% and, therefore, the misclassification is 2 points. So, we can expect two sampling periods of error in the time estimation (2 if only SVM is considered and 3 if only the Bayes/Parzen estimator is taken into account). Fourth, for the H/L and reasoning in the same way, we can expect statistically 7 sampling periods of error, to be compared with 12 and 9 for individual SVM and Bayes/Parzen, respectively. Hence, due to the high success rates achieved for the L to H transitions, one can expect only a statistical reduction of 1 sample around the L to H transition. In the other case, the statistical reduction can be up to 5 samples around the H to L transition. In the case of smooth transitions (figure 10), 5 samples imply an improvement of 50 ms in the determination of the transition time (assuming a sampling period of 10 ms as in this paper).

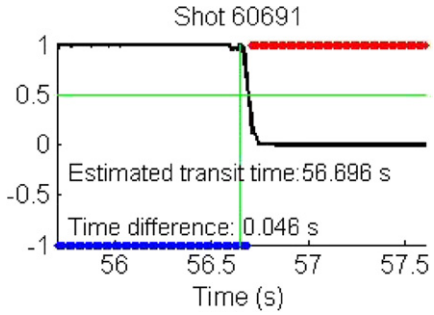

Figure 10. Smooth H to L transition. The meaning of the curves is the same as in figures 8 and 9. Feature vectors are sampled with a period of 10 ms.

The development of training sets that are representative enough of the respective transition is a tedious and very time consuming procedure because it requires a big expert effort to identify transition points. In addition to this, the unavoidable uncertainties that can appear in the identification process should be noted. To this end, future work could be focused on developing iterative techniques to determine the transition times with high precision. The objective would be to start with a reduced training set with only approximated transition times and to converge to a proper solution. Of course, this solution would be representative only of the particular training set. By including new training discharges and repeating the iterative process, a very general model can be developed.

The present technique can be applied not only for off-line analysis (for example, ITER forecasting with a global database of different fusion devices) but also under real-time requirements (control purposes). The training process can demand a significant computational effort but the classification tasks can be carried out very fast.

Acknowledgments

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] Vega J. 2008 *Fusion Eng. Des.* **83** 382–6
[2] Nakanishi H., Hochin T. and Kojima M. 2004 *Fusion Eng. Des.* **71** 189–93

[3] Vega J., Murari A., Pereira A., Portas A., Rattá G.A., Castro R. and JET-EFDA Contributors 2009 *Fusion Eng. Des.* **84** 1916–19
[4] Murari A., Vagliasindi G., Zedda M.K., Felton R., Sammon C., Fortuna L. and Arena P. 2006 *IEEE Trans. Plasma Sci.* **34** 1013–20
[5] Lukianitsa A.A., Zhdanov F.M. and Zaitsev F.S. 2008 *Plasma Phys. Control. Fusion.* **50** 065013
[6] Meakins A.J. 2008 A study of the L–H transition in tokamak fusion experiments *PhD Thesis* Imperial College London, UK
[7] Vapnik V. 1999 *The Nature of Statistical Learning Theory* 2nd edn (Berlin: Springer)
[8] Duda R.O., Hart P.E. and Stork D.G. 2001 *Pattern Classification* 2nd edn (New York: Wiley-Interscience)
[9] Kuncheva L.I. 2004 *Combining Pattern Classifiers: Methods and Algorithms* (New York: Wiley)
[10] Tellaeche A., Burgos-Artizzu X.P., Pajares G. and Ribeiro A. 2007 IEEE *Int. Symp. on Intelligent Signal Processing (WISP) (Alcalá de Henares, Madrid, Spain, 3–5 October 2007)* pp 425–30 http://www.wisp2007.com *Conference Proceedings Book* IEEE Catalog Number 07EX1620, ISBN 1-4244-0829-6 (CD_ROM)
[11] Zimmermann H.J. 1991 Fuzzy set theory and its applications (London: Kluwer Academic Publishers)
[12] 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
[13] Breiman L., Friedman J.H., Olshen R.A. and Stone C.J. 1993 *Classification and Regression Trees* (Belmont, CA/New York: Wadsworth Inc./Chapman and Hall)
[14] Cherkassky V. and Mulier F. 1998 *Learning from Data* (New York: Wiley)
[15] Martinez W.L. and Martinez A.R. 2002 *Computational Statistics Handbook with MATLAB* (London/Boca Raton, FL: Chapman and Hall/CRC Press)