A classification scheme for edge-localized modes based on their probability distributions[a)]

A.Shabbir,[1,2,b)] G.Hornung,[1] G.Verdoolaege,[1,3] and JET contributors[c)]
(EUROfusion Consortium, JET, Culham Science Centre, Abingdon, OX14 3DB, UK)
[1)] *Department of Applied Physics, Ghent University, B-9000 Ghent, Belgium*
[2)] *Max Planck Institute for Plasma Physics, D-85748 Garching, Germany*
[3)] *Laboratory for Plasma Physics, Royal Military Academy, B-1000 Brussels, Belgium*

(Dated: 13 June 2016)

We present here an automated classification scheme which is particularly well suited to scenarios where the parameters have significant uncertainties or are stochastic quantities. To this end, the parameters are modeled with probability distributions in a metric space and classification is conducted using the notion of nearest neighbors. The presented framework is then applied to the classification of type I and type III edge-localized modes (ELMs) from a set of carbon-wall plasmas at JET. This provides a fast, standardized classification of ELM types which is expected to significantly reduce the effort of ELM experts in identifying ELM types. Further, the classification scheme is general and can be applied to various other plasma phenomena as well.

I. INTRODUCTION

High confinement regimes in tokamak plasmas are accompanied by a repetitive magnetohydrodynamic instability of the plasma edge, called the *edge-localized modes* (ELMs)[1]. On the one hand, they are beneficial as they contribute towards impurity control. On the other hand they degrade confinement and large unmitigated ELMs are expected to cause intolerable heat loads on the plasma-facing components (PFCs) in the next-step fusion device ITER.

A first characterization of ELMs is the identification of their type. Hitherto, various types of ELMs have been identified on an empirical and phenomenological basis. In this work, a machine-based classification scheme is developed for the characterization and automatic classification of ELM types, with the aim to distinguish ELM classes (types) in a practical, fast and standardized way.

To this end, two steps are accomplished: ELM feature extraction and classification. Feature extraction involves constructing probability distributions of global plasma parameters and inter-ELM time intervals (also referred to as waiting times) (Δt_{ELM}). Representation through probability distributions allows for an effective treatment of the substantial measurement uncertainties and the inherent stochasticity of ELM properties[2]. In the next stage, we employ the mathematical framework of *information geometry*, which allows a family of probability distributions to be interpreted as a (Riemannian) differentiable manifold[3]. The Fisher information provides a unique metric tensor on such a manifold, which allows for the derivation of geodesics (length-minimizing curves) and the *geodesic distance* (GD) between two points on the manifold[3]. This paves way for the development of a distance-based classifier on the probabilistic manifold. The classifier is then employed for the classification of type I and type III ELMs in an assembled dataset of JET plasmas with PFCs made of carbon fiber composites (hereafter CW).

II. A GEOMETRIC-PROBABILISTIC NEAREST NEIGHBOR CLASSIFIER

The distance-based classification on the probabilistic manifold is performed using the concept of nearest neighbors. The underlying principle of the nearest-neighbor classification is that instances within a dataset will generally exist in close proximity to other instances that have similar properties. In order to classify a test sample (class unknown), the k-nearest-neighbor (kNN) algorithm finds its k closest samples (neighbors) in the d-dimensional training data (class known). The 'closeness' or distance to the training data of a test sample is determined by using a distance metric, such as the Euclidean distance in Euclidean space and the GD on the probabilistic manifold. As illustrated in Fig. 1, the test sample is assigned to the class which is most common amongst its k nearest neighbors. A GD-based kNN classifier offers a number of attractive advantages:

- It makes use of a well-developed mathematical framework for effectively utilizing the information content residing in the distributions of the plasma parameters and ELM properties.
- kNN is non-parametric and does not make any assumptions about the underlying class distribution or the shape of the decision boundary, on the manifold.

[a)] Contributed paper published as part of the Proceedings of the 21st Topical Conference on High-Temperature Plasma Diagnostics, Madison, Wisconsin, June, 2016.

[b)] aqsa.shabbir@ugent.be

[c)] See the Appendix of F. Romanelli et al., Proceedings of the 25th IAEA Fusion Energy Conference 2014, Saint Petersburg, Russia.

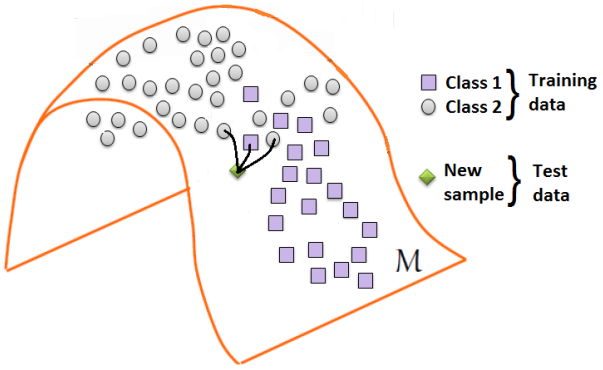

FIG. 1. Illustration of the working of k-nearest neighbor classification on the manifold M. Test sample (probability distribution) is assigned class 2, using 3-nearest neighbor (3NN) classification as class 2 is the majority class amongst its 3 nearest neighbors. The nearest neighbors are ascertained by computing GDs between the test sample (class unknown) and the samples in the training data (class known).

- kNN is intuitively simple. Inference is made directly from the data and there is no model building process on the manifold.

III. CLASSIFICATION OF ELM TYPES

From the JET CW plasmas a dataset spanning over the shot range [50564-76483] and comprising 69 type I, 26 type III and 5 so-called type I high frequency (HF) ELMy plasmas has been constituted. This is essentially the same dataset that has been used earlier for the visualization of the tokamak operational space in[4] and is an extension of the data set used earlier by Webster *et al.* in[5]. The analysis, in this work, has been restricted to time intervals in which the plasma conditions are quasi-stationary with approximately constant heating, gas fueling and central density. Further, all experiments dealing with ELM control and mitigation techniques have been excluded.

TABLE I. Leave-one-out cross-validated (CV) classification success rates (%) using global plasma parameters as predictors and 1-nearest neighbour (1-NN) classifier. Euclidean distance based 1-NN is used for classifying on the basis of the mean (μ) values of plasma parameters and both Euclidean distance based 1-NN and GD-based 1NN are used for classifying on the basis of distributions (μ, σ) of plasma parameters.

Plasma parameters		Distance measure	Leave-one-out CV success (%)		
			I	III	Avg
P_{input}, Γ_{D_2}, B_t	μ	Euclidean	89.2	69.2	84.0
I_p, n_e, δ_{avg}	μ, σ	Euclidean	89.2	69.2	84.0
	μ, σ	**GD**	**95.9**	**84.6**	**93.0**

A. Using global plasma parameters

The global plasma parameters considered for each discharge are: vacuum toroidal field at $R = 2.96$ (B_t) (T), plasma current (I_p) (MA), line-integrated edge density (n_e)(10^{19} m^{-2}), gas fueling (Γ_{D_2}) (10^{22}s^{-1}), input power (P_{input})(MW) and average triangularity (δ_{avg}), where

$$P_{input} = P_{ohmic} + P_{NBI} + P_{ICRH} \tag{1}$$

and

$$\delta_{avg} = \frac{\delta_{lower} + \delta_{upper}}{2}. \tag{2}$$

For simplicity, we assume that the error bars associated with each plasma parameter represent a single standard deviation. Theoretically the underlying probability distribution is Gaussian with the measurement and its error bar constituting the mean (μ) and the standard deviation (σ), respectively.

Table I presents the leave-one-out cross-validated success rates (%) for 1-nearest neighbour (1-NN) classification of type I and type III ELMs using global plasma parameters as predictors. The success rate is defined as the percentage of correct classifications, i.e. the percentage of type I and type III ELMs correctly classified. Class-wise success rates as well as the average classification success rates are presented. It can be noted that the classification using the distributions of the plasma parameters and GD yields significantly higher success rate than that obtained using the Euclidean distance measure or only the mean parameter values as predictors. This demonstrates that the probabilistic description of plasma parameters contains significantly more information than single measurement values (or averages) and that the GD in contrast to Euclidean distance is a more accurate and an intrinsic distance measure for comparing probability distributions.

B. Using ELM waiting times

A robust ELM detection algorithm is used for extracting N ELM waiting times (Δt_{ELM}) from each discharge using the time trace of Balmer alpha radiation from deuterium (D_{alpha}) at JET's inner divertor. Gaussian and 2-parameter (2P) Weibull distributions are then used for modeling the N waiting times extracted from each discharge. Webster *et al.*[5] have recently shown that, based on experimentally motivated assumptions, the 3-parameter (3P) Weibull distribution is a good model for capturing the waiting time statistics. However, a closed form of the GD between 3P Weibull distributions has not been obtained so far. Hence, for ensuring that the developed classification system is computationally efficient and as a first approximation, the 2P Weibull distribution is used. The free parameters of both Gaussian and 2P Weibull distribution are determined using maximum-likelihood estimation and are shown in Fig. 2.

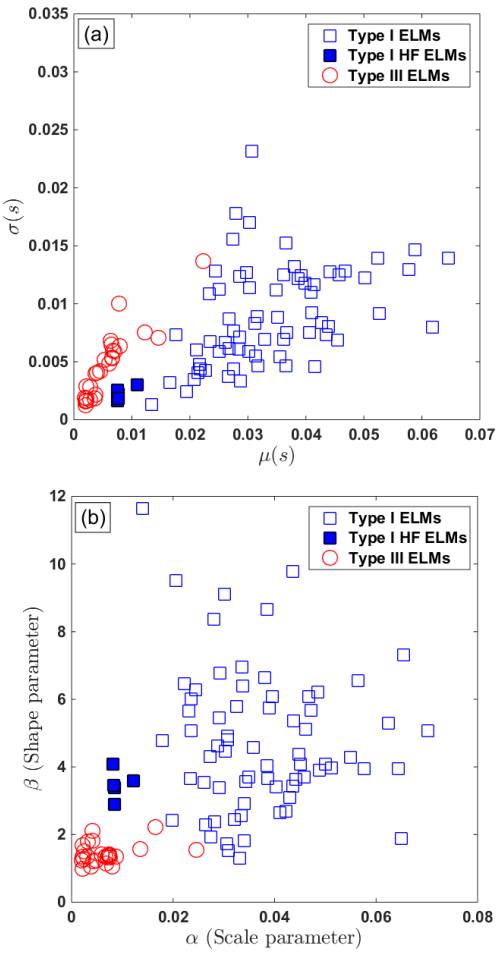

FIG. 2. Maximum-likelihood parameter estimates for (a) Gaussian distribution fit, (b) 2-parameter (2P) Weibull distribution fit to the ELM waiting times (Δt_{ELMs}).

TABLE II. Leave-one-out cross-validated (CV) classification success rates (%) for type I and type III ELMs using mean value and distributions of ELM waiting times as predictors and a 1NN classifier.

Predictors	Distance measure	Leave-one-out CV success (%)		
		I	III	Avg
μ	Euclidean	95.9	84.6	93.0
(μ, σ)	Euclidean	95.9	84.6	93.0
(μ, σ)	**GD**	**97.3**	**96.2**	**97.0**
(β, α)	Euclidean	94.6	80.8	91.0
(β, α)	GD	97.3	92.3	96.0

An examination of Fig. 2, provides various insights. Fig. 2(a) suggests that there is a positive linear correlation between mean and the standard deviation of the waiting times. This implies that type I ELMs, which typically have a higher mean waiting time, tend to have a wider distribution (i.e higher standard deviation) than type III ELMs. Furthermore, both the mean waiting time and its standard deviation appear to be discriminators of ELM type, especially for the discharges which, as far as the distribution of waiting time is concerned, lie at the boundary between type I and type III ELMs. For example, type I HF ELMs have mean waiting times which are smaller than typical type I ELMs but are more similar to type III ELMs. However, they tend to have a smaller standard deviation than the standard deviation of type III ELMs with comparable mean waiting times. Similarly, Fig. 2(b) indicates that β (shape parameter) and α (scale parameter) are together discriminators for ELM type. Type I ELMs typically have a higher value for α than type III ELMs. Also, the information in β appears useful for correctly classifying type I HF ELMs, since they have a higher value of β than the type III ELMs with similar values of α.

This is also reflected in the classification success rates presented in Table II. In consistence with the classification results obtained using global plasma parameters as predictors, GD-based 1NN classification using complete distributions of ELM waiting times yields the highest classification success rate.

IV. CONCLUSIONS AND OUTLOOK

In this paper, a practical, high-accuracy, standardized and automatic classification scheme for ELM types has been presented which can considerably reduce the effort of ELM experts in identifying ELM types. This work clearly elucidates that distributions of plasma parameters contain more useful information than the mere average values. An effective exploitation of this additional information using information geometry results in a superior performance of the classification system. Lastly, the presented classification scheme is generic and can also be applied to other classification problems in fusion plasmas.

The future work will involve applying the presented scheme for classifying additional ELM types, such as type II ELMs as well as for constructing a machine-independent classifier of ELM types.

ACKNOWLEDGMENTS

This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the EURATOM research and training programme 2014-2018 under grant agreement No 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

[1]H. Zohm, Plasma Phys. Control. Fusion, **38** 2 (1996).
[2]A.Shabbir *et al.* IEEE T Plasma Sci. **43** 12 (2015).
[3]G.Verdoolaege *et al.* Nucl. Fusion, **55** 113019 (2015).
[4]A. Shabbir *et al.* Rev. Sci. Instrum. **85**, 11E819 (2014).
[5]A.J. Webster *et al.* Phys. Rev. Lett. **110**, 155004 (2013).