EFDA–JET–CP(14)05/12

A. Pereira, J. Vega, R. Moreno, S. Dormido-Canto,
G. Rattá and JET EFDA contributors

# Feature Selection for Disruption Prediction from Scratch in JET by using Genetic Algorithms and Probabilistic Predictors

# Feature Selection for Disruption Prediction from Scratch in JET by using Genetic Algorithms and Probabilistic Predictors

A. Pereira[1], J. Vega[1], R. Moreno[1], S. Dormido-Canto[2],
G. Rattá[1] and JET EFDA contributors*

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

*[1]Laboratorio Nacional de Fusión. CIEMAT, Madrid, Spain*
*[2]Dpto. Informática y Automática - UNED, Madrid, Spain*
** See annex of F. Romanelli et al, "Overview of JET Results", (24th IAEA Fusion Energy Conference, San Diego, USA (2012)).*

Preprint of Paper to be submitted for publication in Proceedings of the
28th Symposium on Fusion Technology SOFT, San Sebastian, Spain
29th September 2014 – 3rd October 2014

**ABSTRACT**

Recently, a probabilistic classifier has been developed at JET to be used as predictor from scratch. It has been applied to a database of 1237 JET ITER-like wall discharges (of which 201 disrupted) with good results: success rate of 94% and false alarm rate of 4.21%. A combinatorial analysis between 14 features to ensure the selection of the best ones to achieve good enough results in terms of success rate and false alarm rate was performed. All possible combinations with a number of features between 2 and 7 were tested and 9893 different predictors were analyzed. An important drawback in this analysis was the time required to compute the results that can be estimated in 1731 hours (~2.4 months).

Genetic algorithms (GA) are searching algorithms that simulate the process of natural selection. In this article, the GA and the Venn predictors are combined with the objective not only of finding good enough features within the 14 available ones but also of reducing the computational time requirements.

Five different performance metrics as measures of the GA fitness function have been evaluated. The measure F1-score needed 15 generations to reach the highest fitness value, equivalent to assess 420 predictors at 73.5 hours. Accuracy-rate measure required 12 generations (336 predictors at 58.8 hours). Matthew's correlation coefficient (MCC) found the most relevant features after 8 generations (224 predictors, 39.2 hours). Markedness measurement scored 7 generations (196 predictors, 34.3 hours) and the best assessment was the measure called Informedness (the difference of success rate and false alarms), with just 6 generations (168 predictors at 29.4 hours). In all cases, the results show a success rate of 94% and a false alarm rate of 4.21%.

## 1. INTRODUCTION

A disruption is a dramatic extinction of the plasma current with a sudden loss of confinement in a very short time scale. It can produce large forces, strong loads and irreversible damage to the fusion device and surrounding components. Disruption avoidance is extremely important in Tokamak devices. In JET, machine learning methods have been used for the development of automatic systems able to predict incoming disruptions [1]. The objective of a disruption predictor is to predict as early as possible the plasma behavior like disruptive while a discharge is in execution. The Advance Predictor Of DISruptions (APODIS) was developed as a data-driven model based on a combination of several Support Vector Machine (SVM) classifiers. It has been installed in the JET real-time data network [2] and has been very successful during the last metal ITER-like wall campaigns [3]. An optimized APODIS system to predict from scratch [4] obtained also good prediction rates. From scratch means that there is a lack of information during the first training processes and the predictor has to learn without any knowledge about disruptions from the beginning. It happens when a new machine starts its operation (great relevance thinking of ITER). Continuing with this paradigm, recently, a high learning rate probabilistic predictor (based on Venn machines) has been developed at JET [5] under the 'from scratch' approach. Venn predictors are used as probabilistic classifiers

instead of using bare classifiers like SVM. They have the advantage of providing a confidence level for each individual prediction. Moreover, in the specific implementation carried out in [5], it has been possible to reduce the input sample space by means of grouping the data to use a nearest centroid taxonomy. This results in a faster predictor when testing new discharges. Success rate of 94% and false alarm rate of 4.21% were achieved from a database of 1237 discharges, of which 201 disrupted, belonging to the metal wall campaigns at JET. This implementation uses seven signals to characterize the disruptive/non-disruptive plasma state. These signals are processed using 32ms time windows with a sampling frequency of 1 kHz. Two features per signal are computed during the 32ms time windows: mean value and standard deviation of the Fourier Spectrum (after removing the DC component). To ensure the selection of the best features between the 14 available ones, a combinatorial analysis was performed. All possible combinations with a number of features between 2 and 7 were tested and 9893 different predictors were analyzed. It was a costly process and the main drawback was the time required to calculate these results. The computational time was 1731 hours (equivalent to 2.4 months). In this work and with the aim of reducing this time as much as possible, a feature selection method based on GA and the Venn predictors are combined. Moreover, different performance measures are evaluated, demonstrating that is really significant the correct selection of the metric to get quick and successful results. The paper is organized in 5 sections. Section 2 explains several metrics to assess learning algorithms that can be used as fitness function on GA. Section 3 describes the genetic search procedure as an important tool to find impactful variables. Section 4 analyses the results of the five performance metrics used with GA and finally, section 5 summarizes the most relevant contributions in the present paper.

## 2. PERFORMANCE MEASURE

A classifier is, typically, evaluated by a confusion matrix as illustrated in Fig.1a. The columns are the actual class and the rows are the predicted class. TN is the number of negative examples correctly classified (also called true negatives or non-disruptive discharges in the present case). FP is the number of negative examples incorrectly classified as positive (false positives or false alarms), FN is the number of positive examples incorrectly classified as negative (false negatives or missed alarms) and TP is the number of positive examples correctly classified (true positives or disruptive discharges). Therefore, the predictive accuracy can be defined as follows.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

But, Accuracy might not be appropriate when test data are imbalanced and/or the costs of different errors vary markedly [6]. The Receiver Operating Characteristic (ROC) curve is a standard technique for summarizing classifier performance over a range of tradeoffs between TP and FN error rate. On a ROC curve, the x-axis represents the FP rate or false alarm rate, i.e. $FA = FP/(TN+FP) = 1-TN/(TN+FP)$ = 1- specificity and the y-axis represents the recall or success rate, i.e. $SR = TP/(TP+FN)$,

also called sensitivity. The ideal point on the ROC curve would be (0, 1), that is, all positive examples are classified correctly and no negative examples are misclassified as positive, *SR–FA* = 1, but in practice, in the most cases, it does not happen so. The precision for a class is the positive predictive value, i.e. precision = *TP*/(*TP*+*FP*) and the main goal is to improve the sensitivity without hurting the precision, but this objective can be often conflicting, since when increasing the TP for the minority class, the FP for the majority class can also be increased (Fig.1b); this will reduce the precision. The *F1–score* metric is one that combines the tradeoffs of *precision* and *sensitivity*.

$$F1\text{-}score = \frac{2TP}{2TP+FP+FN}$$

However, *Accuracy* and *F1–score* can present specific biases, namely that they ignore performance in correctly handling negative examples, i.e. *NPV* = *TN*/(*TN*+*FN*), thus, *Informedness*, *Markedness* and *MCC* (Matthew's Correlation Coefficient) are formulated in [7] as unbiased measures to avoid the bias of sensitivity, precision and Accuracy measures, respectively.

$$Informedness = sensitivity + specificity - 1$$

$$Markedness = precision + NPV - 1$$

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Regardless of their suitability, the above five equations are contrasted in the fourth section as fitness function that can be used properly on GA.

## 3. FEATURE SELECTION AND GENETIC ALGORITHMS

Feature selection, is the process of selecting the most important features, from a large set of them, by eliminating the redundant and irrelevant ones. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Both of them may have negative effects on classification algorithms, increasing computational time and reducing accuracy and recognition rate. GA has demonstrated the effectiveness to identify variables in the data set as important [8], providing good results in limited computational time. On the other hand, to find the best solution implies testing the whole combination of features and it requires high computational costs and a lot of time consuming. This paper presents a fast method using GA for feature extraction. At the present technique, the algorithm starts with an initial set of random solution called population. A random population of 28 individual is generated. Each individual is also known as chromosome and it is formed by 14 genes (or features). The quality of each chromosome is estimated by a classifier. A probabilistic classifier based on Venn predictors [5] is used. Venn predictors make prediction directly from data by transduction (without generating any rules), instead of repeatedly training classifiers to generate models. Previously, the input sample space (all disruptive/non-disruptive information) is condensed

in a proper way (nearest centroid taxonomy), resulting thus, in a faster and balanced classifier for testing the whole population. At each generation (iteration) of the GA, the 28 chromosomes are evaluated using the fitness function. It plays the most important role in genetic search since offspring for the next generation are determined by the fitness function score, which reflects how optimal the solution is: the higher the number the better the solution. This function has to assess the goodness of the numerical error rates obtained with the classifier. Thus, the output of the classifier is the input of the fitness function and it returns a numerical evaluation representing the goodness of the feature subset. The five performance metrics explained in the previous section were used as measures of the GA fitness function. After that, the best individuals with the best scores are selected for creating the next generation. Two more genetic operators such as crossover and mutation will explore new regions of search space by the combination and the replacing of genes, while keeping some of the current information at the same time. The generational process ends when the termination criterion is satisfied, for instance, the number of generations. Finally, the winner features correspond to the best individual for all generations.

## 4. EVALUATION AND RESULTS

The list of signals is explained in Table 1, which is the same as used in the work [5]. As previously, it consists of 14 features belongs to 7 plasma signals.

The best score reached in terms of success rate and false alarms were 94.00% and 4.21% respectively, and their corresponding features are shown on Fig.2.

At the present work, a random population of 28 individuals (twice the number of features) is generated. By using the same initial population, the genetic search is performed five times, once for each of the five different fitness functions. Prediction from scratch of a single individual with the whole dataset (1237 discharges) lasts 10.5 min; therefore a single generation takes about 4.9 hours.

The evaluation of the five metrics using the combination of GA and Venn predictors is presented on Fig.3. *Informedness* metric matched the highest fitness value with just six generations (168 predictors were assessed in 29.4 hours); on the other hand, the choice of F1-score needed fifteen generations, equivalent to assess 420 predictors in 73.5 hours.

## CONCLUSIONS

Five performance measures were evaluated as GA fitness function. The most relevant characteristics obtained in this analysis are consistent with those achieved previously [5], but with the difference of significantly reducing the employed time (1731 vs. 29.4 hours), an improvement of 98.31%. Informedness, Markedness and MCC show better performance than Accuracy and F1-score metrics, finding in less time the most impactful variables to correctly use on disruption prediction. The first three ones (catalogued as unbiased measurements) are more objectives handling incorrectly classified examples and the measures are better weighted. Therefore, it is really significant the correct selection of the fitness function on GA to get quick and successful results.

**ACKNOWLEDGMENTS**

This work was partially funded by the Spanish Ministry of Economy and Competitiveness under the Projects No ENE2012-38970-C04-01 and ENE2012-38970-C04-03.

This work was supported by EURATOM and 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]. G.A. Rattá, et al., An advanced disruption predictor for JET tested in a simulated real time environment, Nuclear Fusion **50** (2010) 025005

[2]. J.M. López, et al. "Implementation of the disruption predictor APODIS in JET real-time network using the MARTe framework". 18th Real-Time Conference. 11th June – 15th June, 2012. Berkeley, CA (USA). (http://rt2012.lbl.gov/RT2012Abstracts.pdf). Submitted to IEEE Transactions on Nuclear Science

[3]. J. Vega, S. Dormido-Canto, J. M. López, A. Murari et al. "Results of the JET real-time disruption predictor in the ITER-like wall campaigns", Fusion Engineering and Design (2013),

[4]. S. Dormido-Canto, et al. "Development of an efficient real-time disruption predictor from scratch on JET and implications for ITER". Nuclear Fusion. **53** (2013) 113001 (8pp)

[5]. J. Vega et al. Adaptive high learning rate probabilistic disruption predictors from scratch for the next generation of tokamaks. Accepted for publication in Nuclear Fusion.

[6]. N.V. Chawla. Data Mining for Imbalanced Datasets: An Overview. Data Mining and Knowledge Discovery Handbook 2005, pp. **853–867**. ISBN 978-0-387-24435-8

[7]. M.W. David. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies. ISSN: 2229-3981, Vol. **2**, Issue 1, 2011, pp-37–63.

[8]. G.A. Rattá, et al., Improved feature selection based on genetic algorithms for real time disruption prediction on JET. Fusion Engineering and Design. Volume **87**, Issue 9, September 2012, Pages 1670–1678.

| Signal name | Label | Units | Id | Definition |
|---|---|---|---|---|
| Plasma current | Ip | A | 1 | mean(Ip) |
| | | | 2 | std(\|fft(Ip)\|) |
| Mode locked amplitude | ML | T | 3 | mean(ML) |
| | | | 4 | std(\|fft(ML)\|) |
| Plasma internal inductance | LI | | 5 | mean(LI) |
| | | | 6 | std(\|fft(LI)\|) |
| Plasma density | Ne | $m^{-3}$ | 7 | mean(Ne) |
| | | | 8 | std(\|fft(Ne)\|) |
| Diamagnetic energy derivative | dW/dt | W | 9 | mean(dW/dt) |
| | | | 10 | std(\|fft(dW/dt)\|) |
| Radiated power | Pout | W | 11 | mean(Pout) |
| | | | 12 | std(\|fft(Pout)\|) |
| Total input power | Pin | W | 13 | mean(Pin) |
| | | | 14 | std(\|fft(Pin)\|) |

*Table 1: List of signals and features.*

CPS14.795-4c

| Informedness | SR | FA | Features | Generation |
|---|---|---|---|---|
| % | % | % | 1 2 3 4 5 6 7 8 9 10 11 12 13 14 | n |
| 89.79 | 94.00 | 4.21 | 0 1 1 1 0 0 1 1 0 0 1 0 0 0 | 6 |
| 89.79 | 94.00 | 4.21 | 0 1 1 1 0 0 1 1 0 0 0 0 0 0 | 9 |
| 89.79 | 94.00 | 4.21 | 0 1 1 1 1 0 0 0 0 0 1 0 0 0 | 13 |
| 89.79 | 94.00 | 4.21 | 0 1 1 1 0 0 0 0 0 0 1 0 0 0 | 18 |
| 89.79 | 94.00 | 4.21 | 0 1 1 1 1 0 0 0 0 0 0 0 0 0 | 19 |
| 89.69 | 94.00 | 4.31 | 0 1 1 1 0 0 0 0 0 0 1 1 0 0 | 6 |
| 89.69 | 94.00 | 4.31 | 0 1 1 1 0 0 0 1 0 0 0 0 0 0 | 7 |
| 89.69 | 94.00 | 4.31 | 0 1 1 1 0 0 1 1 0 0 0 1 0 0 | 10 |
| 89.69 | 94.00 | 4.31 | 0 1 1 1 0 0 1 1 0 0 1 1 0 0 | 11 |
| 89.69 | 94.00 | 4.31 | 0 1 1 1 1 0 1 1 0 0 1 0 0 0 | 11 |
| 89.69 | 94.00 | 4.31 | 0 1 1 1 0 0 1 0 0 0 1 0 0 0 | 14 |
| 89.69 | 94.00 | 4.31 | 0 1 1 1 1 0 1 0 0 0 1 0 0 0 | 15 |
| 89.69 | 94.00 | 4.31 | 0 1 1 1 0 0 0 1 0 0 1 0 0 0 | 15 |
| 89.69 | 94.00 | 4.31 | 0 1 1 1 1 0 1 1 0 0 0 0 0 0 | 19 |
| 89.69 | 94.00 | 4.31 | 0 1 1 1 0 0 0 0 0 0 0 0 0 0 | 19 |
| 89.69 | 94.00 | 4.31 | 0 1 1 1 1 0 0 0 0 0 1 1 0 0 | 25 |
| 89.69 | 94.00 | 4.31 | 0 1 1 1 1 0 0 1 0 0 1 0 0 0 | 27 |
| 89.69 | 94.00 | 4.31 | 0 1 1 1 0 0 1 0 0 0 0 0 0 0 | 29 |
| 89.69 | 94.00 | 4.31 | 0 1 1 1 1 0 1 0 0 0 0 0 0 0 | 35 |
| 89.69 | 94.00 | 4.31 | 0 1 1 1 1 0 0 1 0 0 0 0 0 0 | 47 |

*Table 2: Features sorted by Informedness.*

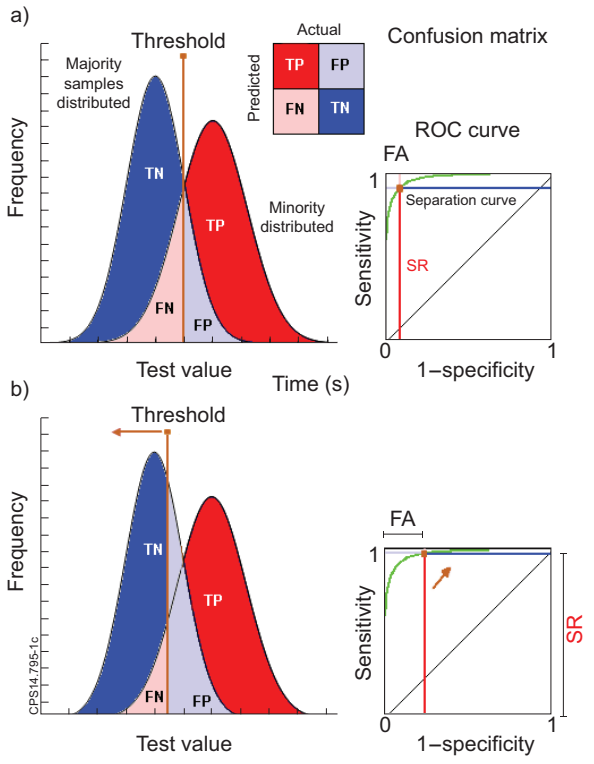
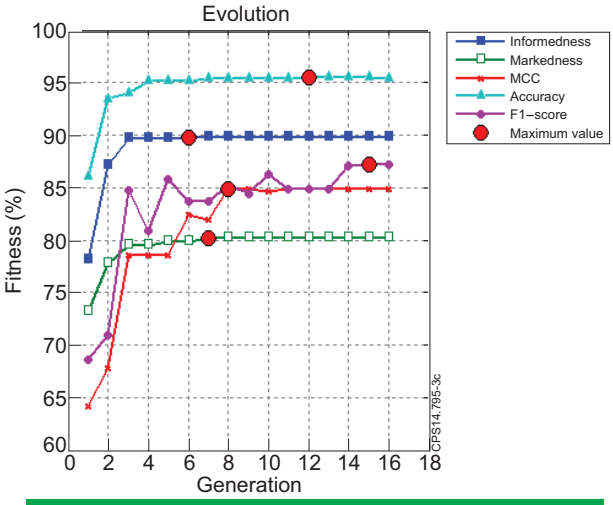



*Figure 1: Useful information to assess learning algorithms.*

| Feature Id. | | | | | | | | | | | | | | SR | FA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | % | % |
| | x | x | | | | | | | | | | | | 94.00 | 4.70 |
| x | | x | | | | | | | | | | | | 92.50 | 5.09 |
| x | x | x | | | | | | | | | | | | 94.00 | 4.31 |
| | x | x | | | | | | | | x | | | | 94.00 | 4.70 |
| x | x | x | x | | | | | | | | | | | 94.00 | 4.21 |
| x | x | x | | | | | | | | x | | | | 94.00 | 4.21 |
| x | x | x | x | | | | | | | x | | | | 94.00 | 4.21 |
| x | x | x | | | | x | x | | | | | | | 94.00 | 4.21 |
| x | x | x | | | | x | x | | | x | | | | 94.00 | 4.21 |
| x | x | x | x | | | | | | | x | x | | | 94.00 | 4.31 |
| x | x | x | x | | | x | x | | | x | | | | 94.00 | 4.31 |
| x | x | x | | | | x | x | | | x | x | | | 94.00 | 4.31 |

CPS14.795-2c

*Figure 2: Features with the best scores reached in a previous work (reference [5]).*




| Metric | Evaluated predictors | Elapsed time (hours) | Generations |
|---|---|---|---|
| Informedness | 168 | 29.4 | 6 |
| Markedness | 196 | 34.3 | 7 |
| MCC | 224 | 39.2 | 8 |
| Accuracy | 336 | 58.8 | 12 |
| F1−score | 420 | 73.5 | 15 |

*Figure 3: GA evolution using five different metrics. Bigger red dots represent the first generation with the maximum fitness value on every evolution.*

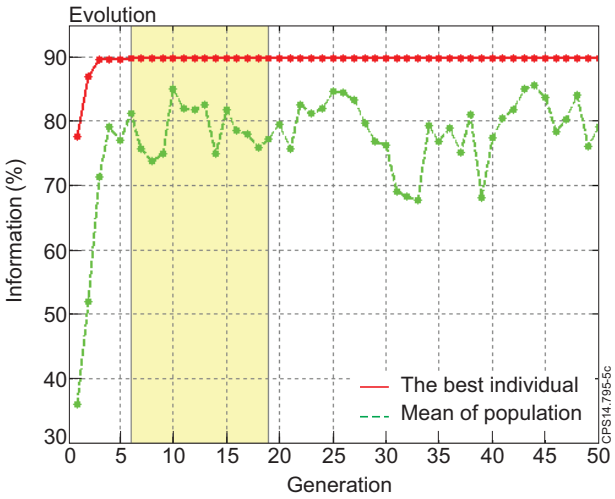



*Figure 4: Yellow region denotes where first emerged all individual with the most relevant features.*