EFDA–JET–CP(08)06-06

J. Vega, G.A. Rattá, P. Castro, A. Murari and JET EFDA contributors

Development of Learning Systems with Data Tours Techniques for Fusion Databases

Development of Learning Systems with Data Tours Techniques for Fusion Databases

J. Vega[1], G. A. Ratt·[1], P. Castro[1], A. Murari[2] and JET EFDA contributors*

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

[1]Asociación EURATOM-CIEMAT, Avenida Complutense 22, E-28040 Madrid, Spain
[2]Consorzio RFX-Associazione EURATOM ENEA per la Fusione. I-35127 Padova, Italy
** See annex of M.L. Watkins et al, "Overview of JET Results ",*
(Proc. 21 st IAEA Fusion Energy Conference, Chengdu, China (2006)).

Preprint of Paper to be submitted for publication in Proceedings of the
8th International FLINS Conference on Computer Intelligence in Decision and Control, Madrid, Spain.
(21st September 2008 - 24th September 2008)

ABSTRACT

Learning systems in massive databases are key elements for the development of both efficient data retrieval methods and data-driven theories. Typically in fusion, waveforms have a very high dimensionality and the construction of classification systems (mainly through unsupervised techniques) without the help of visual techniques is very difficult. The Grand Tour is a 2D visual exploratory method that can be used in the clustering process. Applications to data retrieval and disruption classification in JET are presented.

1. INTRODUCTION

JET is the biggest fusion device in the world and is located in Culham, near Oxford (UK). Diagnostics are measurement systems that transform their observations into electrical signals to be digitized and stored in several formats (mainly waveforms and images). The JET database contains more than 40 Tbytes of data and, nowadays, an amount between 5 and 10 Gbytes/discharge are acquired. The growth rate of the database roughly follows a Moore's Law-like doubling every 2 years. Enhancements to JET in 2007, 2008 and 2009 should result in a maximum of about 60 Gbytes per pulse being collected by 2010.

However, this is small by comparison to the next step machine, ITER. ITER will acquire between 500000 and 1000000 signals/discharge under long pulse conditions (1000s). The ITER database is expected to take about Pbytes/year. Obviously, in both databases, all the data are of no value without mechanisms to efficiently and effectively extract information and knowledge from them.

The use of pattern recognition and data mining techniques provide a big potential to help in the analysis of the massive fusion databases. In this respect, classification systems are crucial elements that stand out over the rest.

1.1 INTELLIGENT DATA ACCESS

Due to the complex and non-linear interactions that are present in thermonuclear plasmas, the signals recorded are often a blend of several underlying processes, mixed together in intricate ways, and generally affected by noise. In addition, some plasma behaviors only appear in an intermittent way as a consequence of instabilities or particular plasma conditions. A breakthrough in data retrieval is the development of techniques to search for data according to technical and scientific criteria instead of time interval or pulse number. Recently, a new model for data retrieval has been proposed and successfully implemented at JET [1]. According to this technique, data access is carried out by using a morphological pattern (structural shape of a signal) as input, whereas the output is the shot numbers and the locations where similar patterns appear.

Taking into account the large size of fusion databases, the search of morphological patterns must be carried out in an intelligent way. This means the entire database cannot be traversed during the searching process. Instead, it is necessary to reduce the searching space to just the signals most likely to contain the required pattern. To this end, classification systems must be developed. Signals

are gathered into groups (clusters) so that the signals within a group are somehow similar. The resulting groups are somehow different and each one represents the reduced searching space to look for similar patterns.

1.2 MODELS OF DATA

Science and engineering are based on the use of first-principle models to describe physical systems. Such an approach starts with a basic scientific model (e.g., Maxwell's theory of electromagnetism) and then builds upon it various applications in electrical engineering. With this approach, experimental data (measurements) are used to verify the underlying first-principle models and to estimate some of the model parameters that are difficult to measure directly. However, in many applications the underlying first principles are unknown or the systems under study are too complex to be mathematically described (as it happens in fusion). Fortunately, fusion devices generate a great amount of information and it is possible to develop models from data.

Within this context, clustering arises as a remarkably rich conceptual and algorithmic framework for data analysis and interpretation. In short, clustering is about discovering structure in collections of data. This leads naturally also to classification. Classifiers are constructs (algorithms) that discriminate between classes of patterns. The discrimination happens through the use of a certain classification boundary, realized by the classifier. The form of the classifier depends not only on the nature of the classifier itself (for instance linear or nonlinear) but also on the values of its parameters. By changing the location of the decision boundary in order to minimize the resulting classification error, the quality of the classifier is affected. The minimization of this error through optimization of the classifier parameters is achieved through learning. In other words, learning can be seen as the process of optimizing the classifier.

2. DATA TOURS

The term 'Exploratory Data Analysis' (EDA) comprises a broad set of graphical tools to visualize and summarize large and complex data sets before making model assumptions to generate hypotheses. However, high data dimensionality is an issue and, therefore, dimensionality reduction is needed. This reduction consists of finding a suitable lower-dimensional space in which to represent the original data. Generally, data are visualized using 2D or 3D scatterplots trying to show interesting structure. Nevertheless, often there is an almost infinite number of possibilities and, for this reason, some methods try touring through the space, looking at many of these representations. These methods are known as data tours. One of them is the Grand Tour [2]. The aim is to look at the data from all possible viewpoints to get an idea of the overall distribution of the n-dimensional data. The Grand Tour provides an overview or tour of a high-dimensional space by visualizing a sequence of 2D scatterplots, in such a way that the tour is representative of all projections of the data. The tour can be halted when an interesting structure is found.

3. GRAND TOUR FOR THE DEVELOPMENT OF LEARNING SYSTEMS

The JET database manages a big amount of data and each one of the stored waveforms can have a typical dimensionality of $10^5 - 10^6$ samples. Giving these quantities, intelligent data retrieval methods are essential tools for data analysis. One specific development is a recognition system of similar waveforms. The input is an entire waveform and the system returns the most similar waveforms ordered by a similarity factor. The algorithm was explained in [3] and an application to JET is shown in [4]. The crucial element to achieve efficiency in the searching process is a multilayer classification system. The Grand Tour has been used to generate the second layer as presented below.

On the other hand, studies about disruptions need the creation of classification systems to categorize the discharges as disruptive or non-disruptive shots. To this end, the Grand Tour can help to create learning systems for proper classification.

3.1 A MULTILAYER CLASSIFICATION SYSTEM FOR THE RECOGNITION OF SIMILAR WAVEFORMS

JET can collect thousands of waveforms/discharge and hence, the database is made up of thousands of signal collections. A signal collection is the complete set of recorded signals belonging to an individual waveform (for example plasma current or diamagnetic energy). The recognition system is based on the creation of a classification system for each signal collection. The waveforms of a signal collection (actually their feature vectors) are classified into a series of categories (clusters). This classification process tries to achieve a convenient set of groups, with a suitable number of waveforms in each cluster, in order to reduce the searching space when looking for similar waveforms for an input signal. The input signal is classified into one of the existing clusters and the computation of the similarity factor is carried out between the input feature vector and the feature vectors of only the selected cluster.

The signal grouping is based on a multilayer classification system whose clustering criteria may evolve in a flexible and dynamical way [3]. Individual clusters can be split at any moment to reach an optimal classification. At least, two layers can be considered. The first one divides the collection into clusters that group shots with the same pulse length. Each first layer cluster can be split into several ones according to a structural shape criterion (Figure 1). The figure shows how the cluster refinement produces groups with lesser number of signals. The splitting process is accomplished choosing one of the projections provided by the Grand Tour.

Figure 2 shows other first layer clusters of JET ECE waveforms with longer pulse length. The structural content is richer than the previous case and the Grand Tour is also able to separate the different shapes.

3.2 HELPING IN DISRUPTION ANALYSIS

Disruptions in Tokamaks are sudden losses of confinement that end the discharge. The more abrupt the disruption the more dangerous can be the consequences for the integrity of the device. Therefore,

the study of disruptions is a research field of primary importance.

Different learning systems can be developed to classify the different types of disruptions. Typically, this is a problem of high dimensionality due to the fact that different signals are used simultaneously to determine an optimal separating hyperplane between classes.

The Grand Tour projections can be used at the end of the feature extraction process to visualize 2D images of the input space. This can help in the selection of feature extractors to build separable systems. In case of nonseparable situations, the projections can give some clues about the way of optimizing the classifier.

As an example, disruptions in JET have been considered. Their feature vectors contain data from 13 signals which have been considered the most important to represent the instabilities leading to disruptions [5]. A series of 262 shots with 116 disruptive discharges was chosen. Temporal evolution segments of the 13 signals were selected as feature vectors. Figure 3 shows separable and nonseparable cases. The former is represented by the left plot where the temporal segment was 100 ms previous to the disruption. The latter (right plot) provides a Grand Tour projection whose temporal segment is 500ms before the disruption.

REFERENCES

[1]. J. Vega. "Intelligent methods for data retrieval in fusion databases". Fusion Engineering and Design. **83** (2008) 382-386

[2]. W.L. Martinez, A.R. Martinez. "Exploratory Data Analysis with MATLAB". Chapman & Hall/CRC. 2004.

[3]. J. Vega, A. Pereira, A. Portas et al. Fusion Engineering and Design. **83** (2008) 132-139.

[4]. J. Vega, G.A. Rattá, A. Murari et al. "Recent results on structural pattern recognition for Fusion massive databases". Proc. of the IEEE International Symposium on Intelligent Signal Processing. ISBN: 1-4244-0829-6 (2007) 949-954.

[5]. A. Murari, G. Vagliasindi, P. Arena et al. Nuclear Fusion **48** (2008) 035010.

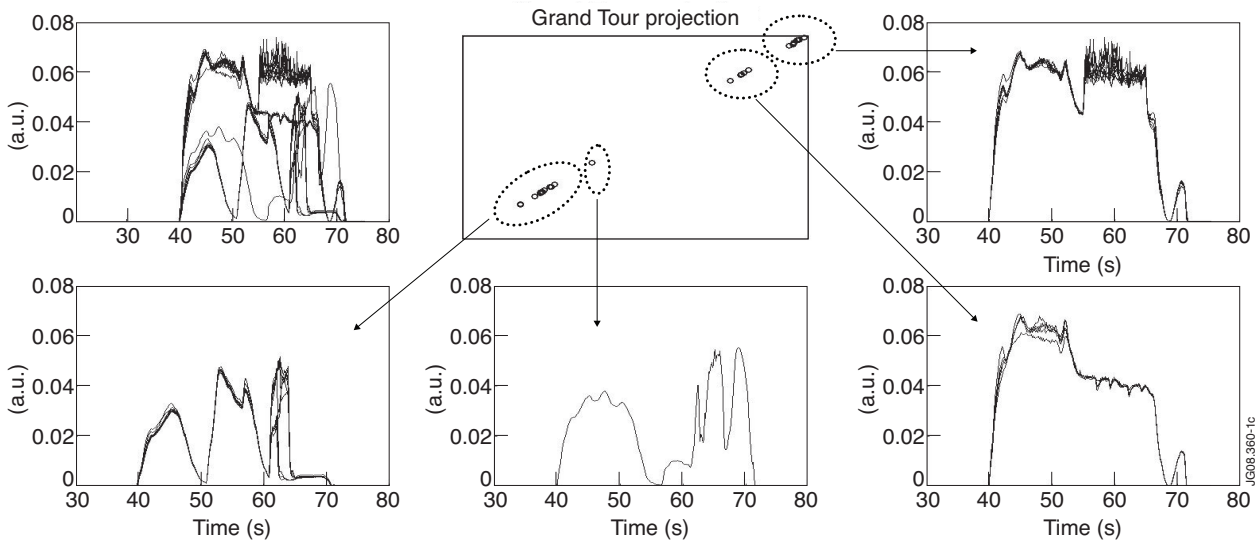

Figure 1: ECE signals from JET database. The top-left plot shows a first layer cluster (all shots have the same length). By choosing a Grand Tour projection, the initial cluster is split into four clearly distinguishable regions. The Grand Tour therefore provides useful indications about how to cluster the data according to the structural shapes of the waveforms.

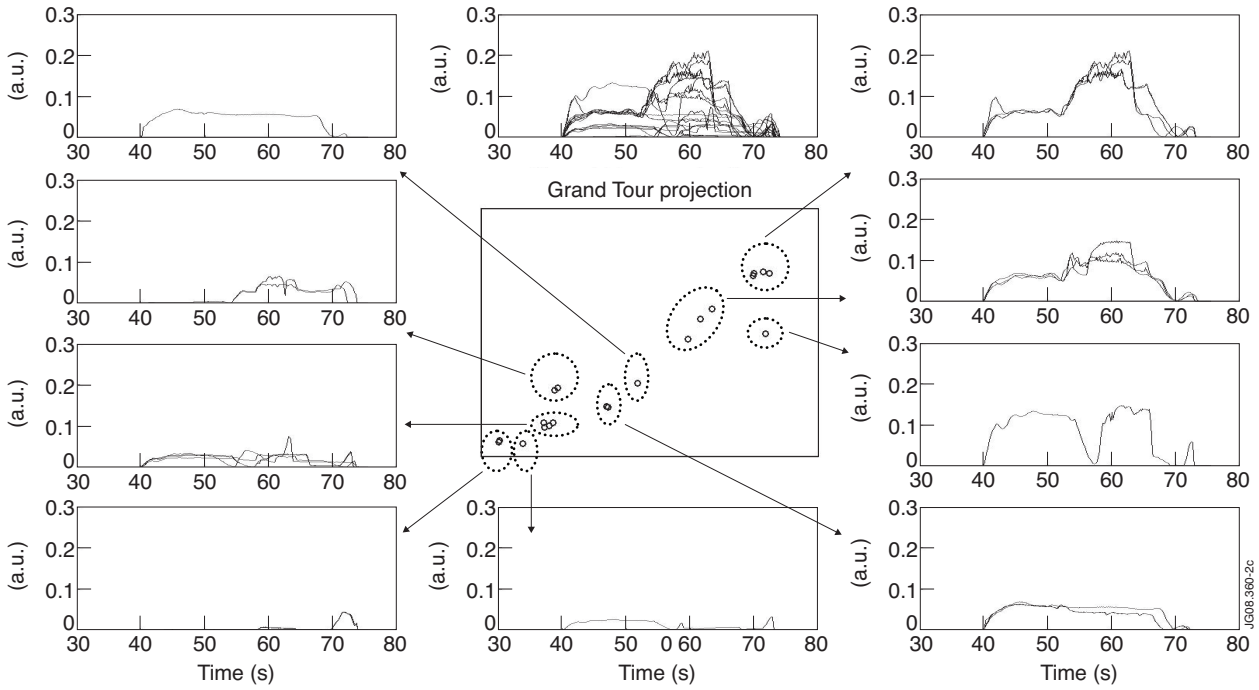

Figure 2: The Grand Tour can distinguish a high number of different structural shapes present in a cluster.

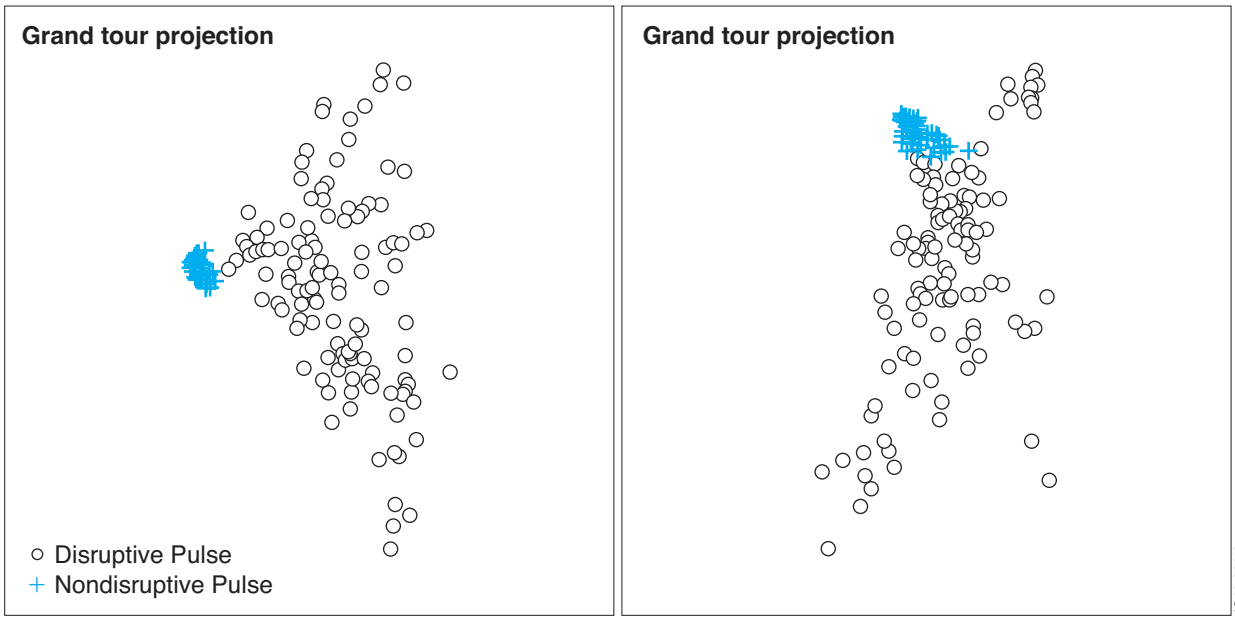

Figure 3: Grand Tour projections showing separable (left) and nonseparable (right) cases for JET disruption classification.