Artificial Intelligence Review (2022) 55:255–289
https://doi.org/10.1007/s10462-021-10032-0

A systemic approach to classification for knowledge discovery with applications to the identification of boundary equations in complex systems

A. Murari[1] · M. Gelfusa[2] · M. Lungaroni[2] · P. Gaudio[2] · E. Peluso[2]

Published online: 30 June 2021

Abstract

Classification, which means discrimination between examples belonging to different classes, is a fundamental aspect of most scientific and engineering activities. Machine Learning (ML) tools have proved to be very performing in this task, in the sense that they can achieve very high success rates. However, both "realism" and interpretability of their models are low, leading to modest increases of knowledge and limited applicability, particularly in applications related to nonlinear and complex systems. In this paper, a methodology is described, which, by applying ML tools directly to the data, allows formulating new scientific models that describe the actual "physics" determining the boundary between the classes. The proposed technique consists of a stack of different ML tools, each one applied to a specific subtask of the scientific analysis; all together they form a system, which combines all the major strands of machine learning, from rule based classifiers and Bayesian statistics to genetic programming and symbolic manipulation. To take into account the error bars of the measurements generating the data, an essential aspect of scientific inference, the novel concept of the Geodesic Distance on Gaussian manifolds is adopted. The properties of the methodology have been investigated with a series of systematic numerical tests for different types of classification problems. The potential of the approach to handle real data has been tested with various experimental databases, built using measurements collected in the investigations of complex systems. The obtained results indicate that the proposed method permits to find physically meaningful mathematical equations, which reflect the actual phenomena under study. The developed techniques therefore constitute a very useful information processing system to bridge the gap between data, machine learning models and scientific theories.

Keywords Machine learning tools · Data driven theory · Support vector machines · Symbolic regression · CART · Knowledge discovery · Boundary equations · Complex systems

A. Murari and M. Gelfusa authors contributed equally to the paper.

✉ M. Lungaroni
michele.lungaroni@uniroma2.it

Extended author information available on the last page of the article

1 Knowledge Discovery in the natural sciences with particular attention to complex systems

Nowadays many fields of science investigate problems of extremely high complexity. Therefore the traditional approach of modelling phenomena starting from first principles is increasingly impractical, when not utterly impossible. Magnetic confinement thermonuclear fusion, to choose an example form Big Physics, is a case in point. The plasmas to be controlled and studied are so complex that a unified treatment is lacking; scientists have to make recourse to models at various physical scales, ranging from particle to fluid, kinetic etc., which have limited generality and do not reproduce satisfactorily many aspects of the dynamics (Wesson 2004). These traditional difficulties are compounded by the recent explosion in the amount of data available. In the last years new sensors and cheap but powerful computing have become commonly deployed in most braches of the experimental sciences, resulting in an authentic data deluge. At CERN, the ATLAS detector has shown the capability of producing Petabytes of data per year. The peak data transmission of the Hubble space telescope was about Gigabytes of data per day. Again in the field of thermonuclear fusion, the data warehouse of the Joint European Torus, the largest Tokamak in operation in the world, is approaching 0.5 Petabytes. Therefore, the inadequacies of theoretical models and the vast amounts of information available have motivated the development of data driven tools, to complement hypothesis driven theories. In this perspective, various machine learning (ML) methods have been refined and to a certain extent applied to the natural sciences. They range from Neural Networks and Support Vector Machines to Fuzzy Logic classifiers; a series of examples from the field of thermonuclear fusion can be found in Murari (2008), Murari (2009), Rattá (2010) Vega (2014). Manifold learning tools, such as Self Organising Maps and Generative Topographic Maps, have provided very good results also in terms of describing the space, in which the relevant physics takes place (Cannas 2013; Murari 2013; Vega 2009; de Vries 2014).

On the other hand, even if the traditional data driven tools are providing quite impressive performance in terms of accuracy, they are scientifically unsatisfactory in many respects. Their main problems in the perspective of applications to the physics of complex systems are: (a) poor “physics fidelity” i.e. excessive discrepancy between the mathematical form of the models and the physical reality of the phenomena investigated (b) insufficient estimates of the uncertainties (c) difficulties to interpret the results in terms of traditional mathematical formulations (d) consequent impossibility to compare the obtained results with mathematical theories based on first principles and (e) lack of extrapolability of the results.

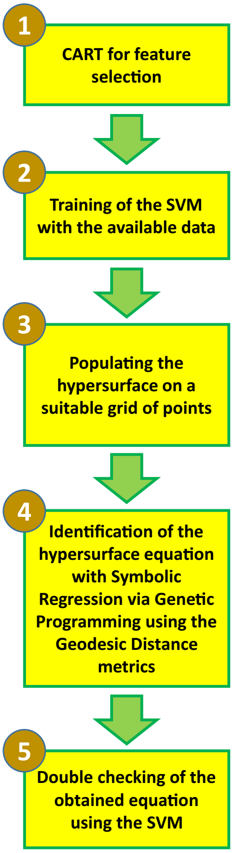
In order to overcome these limitations, a new methodology has been developed to profit from the knowledge acquired by the machine learning tools, but presenting it in a more traditional format, in terms of manageable formulas, which better reflect the reality of the phenomena under study. The techniques, developed in the framework of the activities presented in this paper, address the basic goal of classification. This is a very important task in many scientific applications, both “per se” and as a preliminary step to subsequent investigations.

The goal of classification consists of assigning objects to the most appropriate classes. From pattern recognition to clustering, classification has become an integral part of research in the science of complex systems. Indeed important subjects such as phase transitions, limits of operational space and stability require proper identification of the boundaries in the space of the features. The objective of the analysis presented in this paper is related to deriving mathematical formulas for the boundaries between classes, properly describing the actual physics or chemistry behind the problem. The main idea informing this work resides therefore in combining the learning capabilities of the machine learning tools with the "fidelity" and interpretability of more traditional mathematical formulations, for a more realistic description of the boundaries. It is worth mentioning that the developed tools are conceived for cross sectional data satisfying the i.i.d. assumption, meaning that the data are independently sampled from an identical distribution function. More advanced versions for time series and non-stationary situations are being investigated but are beyond the scope of the present work.

The proposed methodology covers the entire knowledge discovery process, from feature extraction to the final assessment of the quality of the derived models, addressing each fundamental step with a different family of tools. The main motivation behind this choice is that, given the broad and challenging nature of the goal, a stack of individual tools, each one well-tuned to a specific task, is more effective than a single technique. A flow chart of the main steps of the proposed technique, for the case of SVM as main ML tool, is provided in Fig. 1.

The feature extraction phase is performed with a specific refinement of Classification and Regression Trees (CART), the so called noise–based ensembles. The CART approach is particularly useful in this subtask, due to the limited computational burden and the high

Fig. 1 The five main steps of the proposed methodology to identify the best models with a meaningful mathematical form directly from the data


1
CART for feature selection
2
Training of the SVM with the available data
3
Populating the hypersurface on a suitable grid of points
4
Identification of the hypersurface equation with Symbolic Regression via Genetic Programming using the Geodesic Distance metrics
5
Double checking of the obtained equation using the SVM

level of interpretability of the results. Indeed CART rules are easy to interpret and can provide interesting insights about the relevance of candidate features. It is worth pointing out that the effects of the noise and the errors in the measurements are taken into account starting already at this stage, as illustrated in Sect. 2.

The actual classification step is then based on Support Vector Machines (SVM), whose mathematical background is summarised in the Sect. 3, including a probabilistic version very important to quantify the confidence in the results. The choice of SVM is mainly due to their structural stability, their capability to maximize the safety margins in the classification. Given the high accuracy of SVM, the equation of their hypersurface in the original space can be considered an excellent approximation of the boundary between the classes and for this reason they have found many applications in physics (Bahari et al. 2014; Baseer 2018; Beaumont et al. 2011; Clark 2012; Sahin et al. 2016). On the other hand, their mathematical representation of the boundary is extremely non intuitive (see Sect. 3). Indeed referring to systems of the complexity investigated in modern day complex science, the equations of the hypersurface can easily comprise hundreds of support vectors and therefore the equation of the hypersurface contains an equal number of addends. More importantly, the SVM formulation of the boundary equation has typically no relation with the actual physics of the phenomena under study. It has indeed been shown, with many numerical examples (see Sect. 7 and Appendix 1), that the models provided by SVM bear absolutely no resemblance to the ones generating the data. A simple methodology has already been proposed and applied to complex problems, to recover the equation of the boundary in the case of linear kernels (Gaudio 2014). In this paper, a new technique is developed, which is fully general. Indeed the proposed method can be applied to SVM with any type of kernel and even to probabilistic versions; therefore, it has a much wider range of applications than the more traditional techniques. This aspect is very important in the study of complex, nonlinear systems out of equilibrium, which cannot be simply modelled by linear tools or logistic regression.

Symbolic Regression (SR) via Genetic Programming (GP), described in Sect. 4, is the methodology developed to express the output of SVM in an interpretable form appropriate for scientific investigations. Basically, SR via GP is deployed to fit the points on the hypersurface identified, by the SVM as the boundary between the classes. This application of symbolic regression is therefore a development of a technique already very effective for regression (Murari 2012, 2017).

The need for a careful analysis of the uncertainties in the analysis of complex systems has recently emerged as a major topic. In many cases, fundamental sources of uncertainties are the experimental errors. In this perspective, to take into account the error bars of the measurements in a statistically sound way, the formalism of the Geodesic Distance on Gaussian manifolds (GD) has been adopted. Basically this has been inserted in the symbolic regression step: the fitness function of SR via GP is calculated using the GD. The implementation of the Geodesic Distance on Gaussian manifolds is described in Sect. 5.

The combination of the various tools, to derive a physically meaningful equation of the boundary between two classes, is the subject of Sect. 6. In the following Sect. 7 and in Appendix 1, a systematic series of numerical tests are reported, proving the great potential of the proposed methodology. The results of deploying the developed tools for the analysis of experimental databases, addressing completely different phenomena in various scientific disciplines, are reported in Sect. 8. Discussions and lines of future developments are discussed in the last Sect. 9.

Before embarking on the technical description of the developed methodology, a few clarification remarks are appropriate. The approach proposed in this paper is aimed at reconciling the prediction and knowledge discovery capability of machine learning tools with the need to formulate the results in such a way that they can be related to scientific theories and models. It is therefore worth emphasizing that the objective of the present work is not simply improving interpretability of machine learning tools, on which significant work has already been done (Vapnik 2013; Garcia 2009; Vellido 2012; Molnar 2017). The most important aspect indeed is "physics fidelity" i.e. the formulation of the results in mathematical terms, which can be compared with basic theories and models of the various scientific disciplines. The other essential aspect of the proposed methodology, for investigations in the field of complex science, is the principled treatment of the measurement errors, to obtain reliable confidence intervals in the results. This has been achieved with the development of the concepts of Information Technology and in particular the Geodesic Distance on probabilistic manifolds, applied in this work to the task of classification. The other important point to notice is that, as can be seen in Fig. 1, the proposed methodology involves practically all the major fields of machine learning, from rule-based classifiers to Bayesian statistics, genetic programming and symbolic manipulation. Each technique is deployed to solve a specific aspect of the data driven theory process, to which it is particularly suited. The motivation behind the choice of the specific tools, for the applications described in the rest of the paper, will be provided in the sections describing the techniques themselves. On the other hand, it is worth mentioning that the heart of the method is symbolic regression, to formulate the boundary equations in a physically meaningful way. Therefore, other choices of the intermediate steps are perfectly legitimate and would not affect the final results, provided of course their performance are competitive. In any case, this stacked, syncretic approach to knowledge discovery seems to be particularly promising for applications in the natural sciences, in which it is already finding increasing acceptance (see Sect. 8).

2 Noise-based ensembles of CART classifiers for feature selection

Among the rule-based machine learning tools, the so called Classification and Regression Trees (CART) remain among the most developed and widespread. They have been widely implemented for constructing prediction models from data (Breiman 1984; Lungaroni 2018). Such models are derived directly from the available databases by recursively partitioning the feature space and fitting a simple prediction rule at each partition. The final partitioning, once properly optimised, consists therefore of a series of rules that can be represented graphically by a decision tree. The performance of classification trees are typically quantified in terms of misclassification costs. The algorithms of this family search the whole database exhaustively, to determine, for each variable, which value minimizes the total impurity of its child nodes. To quantify the purity of a node, the version of CART implemented in this paper uses a generalization of the binomial variance called the Gini Impurity index.

Decision trees are very practical and easy to interpret but present a significant drawback: their sensitivity to the specific data used for their training. Indeed, a small change in the inputs (for example even using a subset of the training data) can imply a major variation in the resulting decision tree and in turn quite different predictions. To overcome this issue and to increase the success rate of the results, it is typically very advantageous to adopt the approach of ensemble rule-based classifiers, based on the concept of weak learners. A 'weak' learner (either classifier or predictor) is just any machine learning tool, which generates models that perform relatively poorly but are computationally simple (Breiman 1984). The relatively limited computational resources required allow training various versions of such weak learners, which can then be pooled together (via Bagging, Random Forests etc.) to create a "strong" ensemble classifier.

One of the main issues of the measurements in the experimental sciences is often the high levels of noise. This noise is very difficult to reduce; the sources of noise are many and independent. Even if these uncertainties are a potential issue, they suggest a complementary approach to the method of building ensembles of weak classifiers. The idea consists again of collecting ensembles but not with subsets of the original data, as in bagging and random forests; on the contrary the various training sets are obtained by the original one summing random noise to the measurements. The random noise is generated from Gaussian distributions with variance equal to the error bars of the measurements, estimated on the basis of the experimental equipment characteristics. To each realization of the noise corresponds a different weak learner. The number of trees can be increased until the accuracy begins to saturate instead of improving. This approach, called Noise-based Ensemble, can be implemented as a step preliminary to building ensembles of CART trees and allow taking into account the experimental errors from the stage of feature selection.

In the application described in this paper, the Noise-based Ensembles of CART trees are used for feature selection. They are trained to classify, using an extensive set of potential features. So the input space is the set of features and the output the classes. Once CART trees have been trained, a simple inspection of the trees and the rules allow identifying the most relevant features, to be used in the following steps of the methodology. Another alternative consists simply of progressively pruning the trees until the performance becomes unacceptable and keep only the remaining features. An example of the potential of this approach in real life applications, characterised by sparse data and high noise levels, is provided in Sect. 8, while additional background on the technique can be found in (Murari 2019, 2020).

3 Traditional and probabilistic SVM

Traditional SVM are very powerful classifiers, whose principles of operation can be summarised in intuitive terms as follows. In the case of binary classification, such as the practical examples discussed in this paper, SVM map the input examples of the two classes into a high-dimensional space, through a non-linear mapping implemented with the help of suitable kernels (Vapnik 2000). The risk of misclassification is then minimised by identifying an optimal separating hyperplane in this high dimensional feature space. Such minimization of the error risk is achieved by maximizing the margins between the hyperplane and the closest points of each class. These examples closest to the separating hyperplane are called *Support Vectors* (SV). The minimization of the error risk, obtained with the maximization of the margins, is performed in the projected space by minimizing a quadratic functional with appropriate constraints.

SVM therefore basically consist of suitable kernels, which map the inputs into higher dimensional spaces, where the classification becomes a linearly separable problem and can be solved with traditional quadratic programming methods based on the Lagrange multipliers.

In mathematical terms, given a training set of l samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)$, $x_i \in \Re^n$, for a binary classification problem (*i.e.* $y_i \in \{+1, -1\}$), SVM estimates the following decision function:

$$D(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i y_i H\left(\mathbf{x}_i, \mathbf{x}\right) \tag{1}$$

where $H\left(\mathbf{x}_i, \mathbf{x}\right)$ is a kernel function and the parameters $\alpha_i, i = 1, \ldots, \ell$ are the solutions of a quadratic optimization with linear constraints (Vapnik 2000).

At this point a clarification of the terminology is probably appropriate. SVM find a separating hyperplane in the transformed space. On the other hand, the hyperplane is expressed in terms of Support Vectors in the original space, in which the boundary is a hypersurface. Since typically in the study of complex systems, researchers are interested in equations in the original space, and not in the transformed one, the boundary between the two classes will be indicated with the term hypersurface and not hyperplane in the following. Indeed, another advantage of the SVM is that their results are expressed in terms of the inputs in the original space.

The second and third real life examples described in Sect. 8 adopt this approach of the traditional SVM, with a radial basis function kernel and a regularization strength λ equal to 1/*n*, where *n* is the number of observations. On the other hand, the availability of classifiers, which can output a probability, would be extremely useful in most applications typical of complex science. Unfortunately, traditional SVM, as just described, provide only a distance to a hyperplane, in the form reported in Eq. (1). Their basic version has therefore to be extended to associate a probability to the outputs of their classification (Platt 2000; Steinwart 2008). One possible solution consists of reformulating the SVM output in terms of a probability with the Bayes rule, according to the formula:

$$P(y = 1|x) = \frac{p(x|y = 1)P(y = 1)}{\sum_{i=-1,1} p(x|y = i)P(y = i)} \tag{2}$$

In Eq. (2) *y* indicates the label of one of the classes. *P(y = 1)* is the prior probability and *p(x|y = 1)* is the likelihood. Therefore, to convert the outputs of traditional SVM to probabilities, two quantities have to be determined: the prior probability and the likelihood. In many applications, the natural choice of the prior probability is the percentage of examples seen, up to a certain point in time in the experiments or observations, belonging to the class to which the SVM labels the new example.

The most challenging aspect of relation (2) resides in the evaluation of the likelihood. If a solid and reliable estimate of the likelihood is not viable for any reason, theoretical investigations and practical considerations have shown that one advantageous alternative consists of remapping the distance to the hyperplane to a probability by using a sigmoid function (Platt 2000; Steinwart 2008):

$$P(y = 1|d) = \frac{1}{1 + \exp(Ad + B)} \tag{3}$$

In Eq. (3) A and B are two fitting parameters, whereas *d* is the distance of the examples to the SVM hyperplane. Equation (3) therefore allows converting directly the distance to the hyperplane, provided by traditional SVM, into a probability. This conversion takes place after the training; the distances of the examples in the training set are used to fit the parameters of the sigmoid (3). The sigmoid is constrained to be centred on the hyperplane, because points at distance zero from it have equal probability of belonging to any of the two classes. To obtain the points to be fitted with symbolic regression (see next Sect. 4), it is sufficient to select the most appropriate probability threshold (typically the one with better performance in terms of success rate). The points at that level of probability are the inputs to the fitting part of the procedure. This solution of fitting a sigmoid is the one used in the first real life example described in Sect. 8 of the paper.

4 Symbolic regression via genetic programming for physics fidelity

As already briefly discussed in Sect. 1, the objective of the present work is the development of a series of techniques, aimed at converting the models of machine learning classifiers into more realistic mathematical forms. In this context realistic means that the equations of the boundary should reflect the actual phenomena, determining the frontier between the classes. In practice, for the case of SVM classifiers, this task implies expressing the hypersurface separating the classes more realistically than as a sum of hundreds of kernel functions or as a series of points at the same probability (in the case of probabilistic SVM). To this end, the main tool used is Symbolic regression via Genetic Programming. The methods developed on the one hand allow identifying the most appropriate mathematical expression for the hypersurface with minimal "a priori" hypotheses. In this way therefore the potential of SVM is fully exploited and no unnecessary restrictions are imposed on the form of the solutions. On the other hand, the complexity of the obtained solutions can be controlled, allowing to find the best trade-off between complexity, success rate of classification and realism of the final models, depending on the objectives of the study. The rest of this section provides a brief overview of SR via GP for the reader convenience; more details can be found in the references (Schmid 2009; Koza 1992; Murari 2012; Murari 2015).

Fig. 2 The main steps to express a model identified by SVM in more traditional mathematical notation

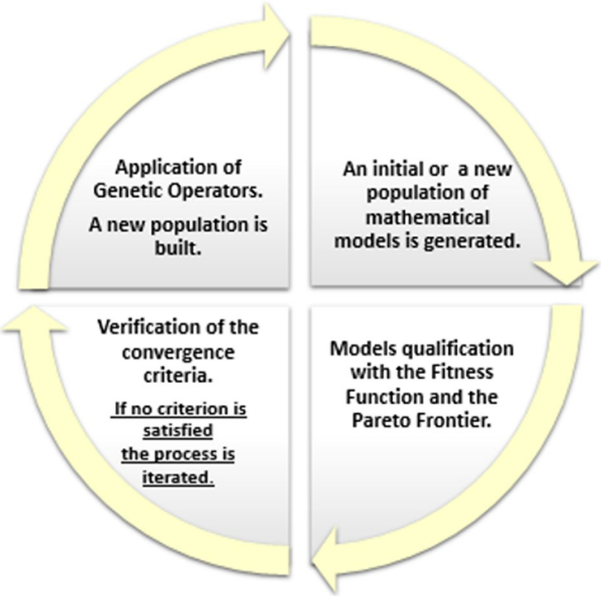

Application of Genetic Operators.
A new population is built.
An initial or a new population of mathematical models is generated.
Verification of the convergence criteria.
If no criterion is satisfied the process is iterated.
Models qualification with the Fitness Function and the Pareto Frontier.

Fig. 3 An example of syntax tree structure for the function $2R+(S/R^2)$The function operator nodes (green) and the variable or constant nodes (red) are reported in different colours

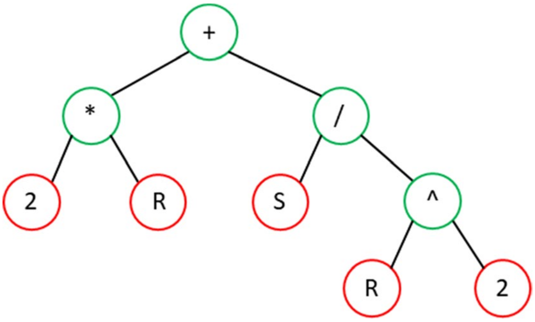


Table 1 Types of function nodes included in the symbolic regression used to derive the results presented in this paper. x_i and x_j are the generic independent variables

Function class	List of functions
Arithmetic	*c (constants),* +,-,*,/
Exponential	*exp(x_i),log(x_i),* *power(x_i, x_j), power(x_i,c)*
Squashing	*logistic(x_i),step(x_i),sign(x_i),* *gauss(x_i),tanh(x_i), erf(x_i),erfc(x_i)*
Trigonometric	*sine, cosine, hyperbolic sine,* *hyperbolic cosine*

The method of SR via GP consists of testing various mathematical expressions to fit a given database. The main steps to perform such a task are reported in Fig. 2. The various candidate formulas are expressed as trees, composed of functions and terminal nodes. A simple example of this form of knowledge representation is provided in Fig. 3. The function nodes can be standard arithmetic operations and/or any mathematical functions, squashing terms as well as user-defined operators (Schmid 2009; Koza 1992). The function nodes, included in the analysis performed in this paper, are reported in Table 1. This representation permits to steer the models towards physics fidelity by proper selecting the basis functions and/or the structure of the trees. Moreover expressing the formulas as trees allows an easy implementation of the next step, symbolic regression with Genetic Programming (GP). Genetic Programs are computational methods able to solve complex optimization problems (Schmid 2009; Koza 1992). They have been inspired by the genetic processes of living organisms. They work with a population of individuals, e.g. mathematical expressions in our case. Each individual represents a possible solution, a potential boundary equation in the present application. An appropriate fitness function (FF) is selected to measure how good an individual is with respect to the database. A higher probability to have descendants is assigned to those individuals with better FF. Therefore, the better the adaptation (the value of the FF) of an individual to a problem, the higher is the probability that its genes are passed to its descendants.

In more detail, the first step of the method is the generation of the initial population of formulas for the boundary between two classes; then the algorithm assesses the quality of each element of the population by evaluating its performance with the metric expressed by the FF. In the following step, as with most evolutionary algorithms, genetic operators (Reproduction, Crossover and Mutation) are applied to individuals that are probabilistically selected on the basis of the FF, in order to generate the new population. This means that better individuals are more likely to have more children than inferior individuals. When a stable and acceptable solution is found or some other stopping condition is met (e.g., a maximum number of generations or acceptable error limits are reached), the algorithm provides the solution with best performance in terms of the FF.

The fitness function is a crucial element of the genetic programming approach and it can be implemented in many ways. To derive the results presented in this paper, the AIC criterion (Akaike Information Criterion) has been adopted (Burnham 2002) for the FF. The classic form of the AIC indicator is:

$$\mathrm{AIC} = 2k + n \cdot \log\left(\frac{\mathrm{RMSE}}{n}\right) \tag{4}$$

In Eq. (4), RMSE is the Root Mean Square Error between the data and the model predictions, k is the number of nodes used for the model and n the number of examples provided, i.e. the number of entries in the database (DB). The AIC, as the other indicators used in this work, is to be minimised: the lower the FF the better the model. The FF parameterized above allows rewarding the goodness of the models, thanks to the RMSE, but at the same time their complexity is penalised by the dependence on the number of nodes.

In practice, quite often the quality of the data and of the prior knowledge are not enough to converge on a single best model. The most common and effective way to handle this situation consists of making recourse to the Pareto Frontier (PF). PF is the collection of the best models, according to the FF, for each level of complexity. The PF derives its name from the usual form of plotting the FF versus complexity; a generic example is shown in Fig. 4. The best points for each level of complexity constitute the Pareto Frontier, which presents typically a shape similar to a letter L. The models around the inflexion point of the PF are the ones achieving the best trade-off between accuracy and complexity and are the ones, to which to apply the techniques discussed in the rest of this section, to converge on the final choice.

To assess the quality of the final models, the well-known criteria of BIC (Bayesian Information Criterion) and Kullback–Leibler (KLD) divergence have been used. The BIC criterion is defined as (Burnham 2002):

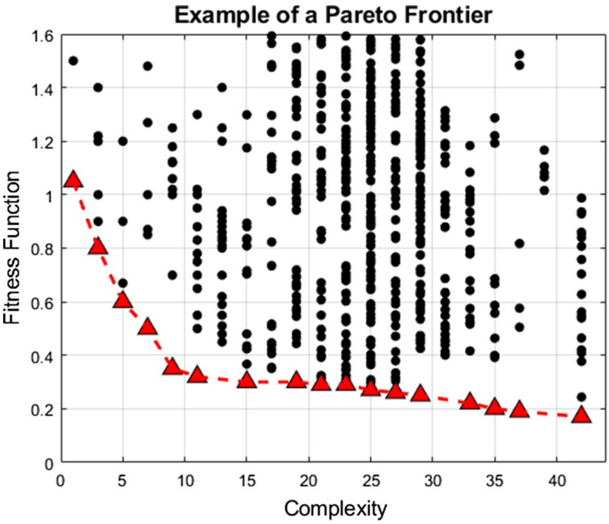


Fig. 4 Typical shape of a generic Pareto Frontier, identified by the red triangles

$$\mathrm{BIC} = n \cdot \log\left(\sigma^2_{(\in)}\right) + k \cdot \log(n) \tag{5}$$

where $\in = y_{data} - y_{model}$ are the residuals, $\sigma^2_{(\in)}$ their variance and the others symbols are defined in analogy with the AIC expression. Again the better the model, the lower its BIC. A more sophisticated form of both the AIC and BIC indicators, to take into account the error bars of the measurements using the formalism of the GD, is introduced in the next section.

The aim of the KLD is to quantify the difference between the computed probability distribution functions, in other words to quantify the information lost when $p\left(\bar{y}_{model}\left(\vec{x}\right)\right)$ is used to approximate $q\left(\bar{y}_{data}\left(\vec{x}\right)\right)$ (Burnham 2002). The KLD is defined as:

$$KLD(P||Q) = \int p(x) \cdot \log\left(\frac{p(x)}{q(x)}\right) dx \tag{6}$$

The Kullback Leibler Divergence assumes positive values and is zero only when the two probability distribution functions (pdfs), p and q, are exactly the same. In our application p is the pdf of the data, considered the reference, and q the pdf of the model estimates. Therefore the smaller the KLD is, the better the model approximates the data, i.e. the less information is lost by representing the data with the model.

The last step of the methodology is the non linear fitting of the models. Once the best mathematical expression to interpret the data has been found with SR via GP, it is necessary to perform a final fitting of the database for two main reasons. First, modern non linear fitting routines allow fine tuning the parameters of the models, improving their generalisation capability. Even more importantly, with this final step it is possible to associate confidence intervals to the parameters of the equations, which is an essential aspect in many scientific applications. A detailed overview of SR via GP for scientific applications is provided in (Murari 2013, 2016).

5 Geodesic distance on Gaussian manifolds to include the effects of the error bars

In this section the geodesic distance on probabilistic manifolds is introduced in subsection 5.1. The use of the geodesic distance in the SR is then detailed in subsection 5.2.

5.1 Geodesic distance

As seen in the previous section, the goal of SR via GP is to extract the most appropriate formulas to describe the available data. To achieve this, typically a quantity somehow proportional to the sum-of-squares of the distances between the data and the model predictions is used in the FF (the RMSE in Eq. 4 and the variance in Eq. 5). In this way, SR is implicitly adopting the Euclidean distance to calculate the (dis)similarity between data points and predictions. However the Euclidean distance implicitly requires considering all data as single infinitely precise values. This assumption can be appropriate in other applications but it is obviously not the case in the science of complex systems, since all the measurements available typically present an error bar. An alternative idea is to use a new distance between data, which would take into account the measurement uncertainties. The additional information provided by this distance is more coherent with the nature of the available features and renders the final results more robust (Murari 2013; Lungaroni 2019).

The principle, behind the approach proposed in this paper, consists of considering the measurements not as points, but as Gaussian distributions. This is a valid assumption in many scientific applications. In many complex systems, in particular, the measurements are often affected by a wide range of noise sources, which from a statistical point of view can be considered random variables. Since the various noises are also typically independent and additive, they can be expected to lead to measurements with a global Gaussian distribution around the most probable value, the actual value of the measured quantity. Each measurement can therefore be modelled as a probability density function of the Gaussian type, determined by its mean μ and its standard deviation σ:

$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{7}$$

Modelling measurements not as punctual values, but as Gaussian distributions, requires defining a distance between Gaussians. The most appropriate definition of distance between Gaussian distributions is the geodesic distance (GD), on the probabilistic manifold containing the data, which can be calculated using the Fischer-Rao metric (Amari 2000). For two univariate Gaussian distributions $p_1\left(x|\mu_1,\sigma_1\right)$ and $p_2\left(x|\mu_2,\sigma_2\right)$, parameterised by their means μ_i and standard deviations $\sigma_i(i = 1, 2)$ the geodesic distance GD is given by:

$$GD(p_1||p_2) = \sqrt{2}\ln\frac{1+\delta}{1-\delta} = 2\sqrt{2}\tanh^{-1}\delta,\ \text{where } \delta = \left[\frac{\left(\mu_1-\mu_2\right)^2 + 2\left(\sigma_1-\sigma_2\right)^2}{\left(\mu_1-\mu_2\right)^2 + 2\left(\sigma_1+\sigma_2\right)^2}\right]^{\frac{1}{2}} \tag{8}$$

The meaning of the GD can be appreciated by inspecting Fig. 5, which reports the distance between two couples of Gaussian distributions. The distance between the means of the members of the two couples is the same. On the other hand, the Gaussian pdfs of

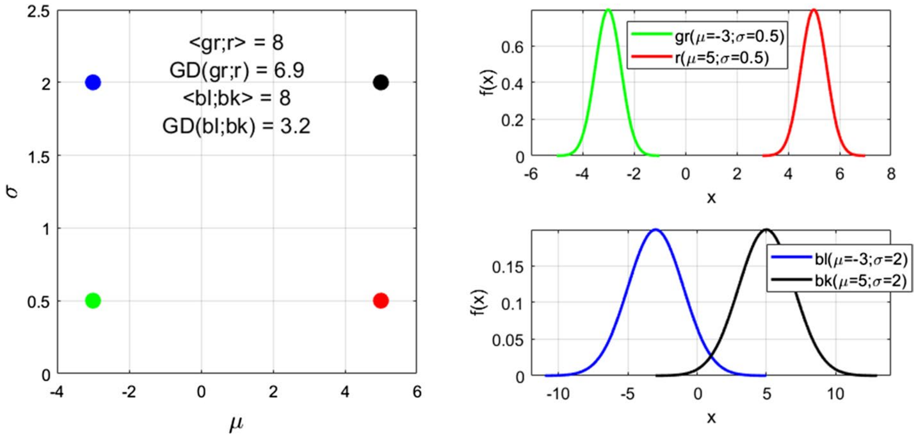


Fig. 5 Examples to illustrate how the GD determines the distance between two Gaussians. The two couples of pdf in the figure have the same mean but different σ. The geodesic distance between the two with higher σ is much smaller. GD indicates the geodesic distance and < > the Euclidean distance

one couple have a standard deviation an order of magnitude higher the other. The distance between the pdfs with higher standard deviation is therefore significantly lower than the one of the more concentrated pdfs, which is intuitively and conceptually correct since they overlap much more. This property of the GD increases the robustness of the results and reduces the risk of overfitting, as verified with a series of numerical tests (see also next subsection).

5.2 Use of geodesic distance in symbolic regression

To take into account the measurement errors in a statistically sound way, the last step required consists of inserting the GD into the SR. To this end, a good solution has been obtained by replacing the RMSE and variance with the GD in the AIC and BIC criteria, according to the following formulas:

$$AIC = 2k + \sum_i GD_i \tag{9}$$

$$BIC = n \cdot \log\left(\sum_i GD_i\right) + k \cdot \log(n) \tag{10}$$

where the symbols have the same meaning as in formulas (8) and (9) and the index i runs over the entries of the database. It is worth pointing out that this idea of inserting the GD in the FF of the SR is another original development proposed by the authors.

Since in the genetic programming, implementing symbolic regression, the GD is to be calculated as the distance between the experimental values and the estimates of the model, the Gaussian parameters μ and σ must be properly chosen. For the experimental data, the typical assumption is to take the measured value as μ, assuming that the average value is the most likely measurement. For the standard deviation, a reasonable assumption is to adopt the value of the error bars in the experimental points. With regard to the model, the point estimate is considered the mean and the confidence intervals are used for σ (Murari 2017). Of course, these estimates are relatively straightforward for the examples discussed in the present paper, because the uncertainties are supposed to obey a Gaussian distribution. In case of different statistics, more sophisticated density estimation technique could be necessary.

In practice, it has been tested with tens of equation that the SR using GD is practically never outperformed by the SR using the RMSE or the variance. Moreover, the GD is also more robust against outliers, as proved for regression in (Murari 2016; Murari 2017). It is indeed a well-known statistical fact that the RMSE and variance are not very robust indicators and are particularly vulnerable to outliers. As an example of these tests, the following equations have been used to generate synthetic data:

$$
\begin{aligned}
f_1 &= \cos\left(x_1 \cdot x_2\right) + \sin x_1^{0.5} \\
f_2 &= \cos\left(\frac{x_1}{x_2}\right) + 2x_3 \cdot \left\{1/\left[1+\exp\left(-0.8\cdot x_2\right)\right]\right\} \\
f_3 &= x_1^{0.8} + \frac{\left\{1/\left[1+\exp\left(-0.6\cdot x_2\right)\right]\right\}}{x_3} \\
f_4 &= x_1 \cdot x_2 \cdot \exp\left(-x_3\right) + 2x_3
\end{aligned}
\tag{11}
$$

The range of variations of the independent variables, for the examples reported in the following, is:

$$
\begin{aligned}
x_1 &= 0.015 \ldots 3.9 \\
x_2 &= 0.044 \ldots 1.97 \\
x_3 &= 0.268 \ldots 2.178
\end{aligned}
\tag{12}
$$

Two different types of noise have been implemented: Gaussian noise, of zero mean and standard deviation equal to a fixed percentage of the mean value of the functions, and noise with outliers. The distribution of the outliers has been modelled with a second Gaussian with a mean different from zero. The weight of this second Gaussian, generating the outliers, can be selected. In general, for tens of different synthetic databases and a number of outliers from 5 to 50% of the entries, SR with the GD outperforms systematically the version using the RMSE. SR with GD manages always to approximate the generating functions not worse than the version with RMSE and it provides better results in about 50% of the cases.

6 Combining SVM and symbolic regression for boundary equations

From a methodological perspective, this section is the heart of the paper; indeed it is meant to explain in detail how SR via GP and the SVM technology can be combined to derive realistic equations of the boundary between classes, suitable to scientific investigations. The subject of Subsection 6.1 is the description of the technique developed to find points on the hypersurface identified by the SVM. Subsection 6.2 explains how symbolic regression can be deployed to fit the previously derived points, in such a way to derive more realistic boundary equations.

6.1 How to find points on the SVM hypersurface

In order to interpret the results produced by the SVM, the first step consists of determining a sufficient number of points on the hypersurface separating the two classes. These points can be then given as inputs to the SR to obtain a more manageable equation for the hypersurface. In the case of probabilistic SVM, obtaining the points on the boundary is technically very simple. The main decision to be taken is the choice of the most appropriate value for the probability threshold to separate the classes; this can be achieved on the basis of the success rate and the objectives of the classification.

Obtaining the hypersurface points in the case of a traditional, non-probabilistic, SVM is a bit more involved and requires a specific procedure described in the rest of this subsection. A mesh is built first, with resolution equal or better than the error bars of the measurements used as inputs to the SVM. The limits of the domain are defined by the ranges of variables. Obviously, more grid points and a better refined mesh can improve robustness and convergence (but not accuracy since the grid is already finer than the error bars of the measurements). Therefore, the total number of grid points can be set based on computational limitations. On the other hand, for selecting the number of intervals more efficiently, a good solution consists of allocating more intervals along the direction of stronger curvature.

After building the grid, the algorithm starts selecting the SVs on the positive side of the hypersurface and moves towards the SVs on the other side, one point on the mesh at the time. At each step, the distance to the hypersurface is computed using the already trained SVM. If the distance remains positive, the process is repeated, since the new point remains on the same side of the hypersurface. When the distance of a new point changes sign, the two points with different signs are considered points on the hypersurface. This assumption is more than reasonable because, by construction of the mesh, these points, for which the distance changes sign, are within a distance from the hypersurface equal or smaller than the error bar of the features (typically measurements). Therefore, for all practical purposes in the science of complex systems, these points are sufficiently close to the hypersurface to be considered on it. This way to obtain SVM hypersurface points for synthetic data is shown pictorially in Fig. 6.

6.2 Deriving the equation of the hypersurface with symbolic regression

The previous subsection has described how to find a sufficient high number of points close to the hypersurface, identified by the SVM. The next phase of the method consists of deriving the equation of the hypersurface itself with the help of SR via GP, using the points obtained with the method detailed in Subsection 6.1. Indeed SR can be directly run to fit the points close to the hypersurface; to maximise the efficiency of this step, it is wise to choose as the independent variable the quantity with the largest dynamic range. Once obtained the best equation identified by SR via GP, what remains is to evaluate the quality of the model. A natural and immediate approach consists of applying the statistical indicators described

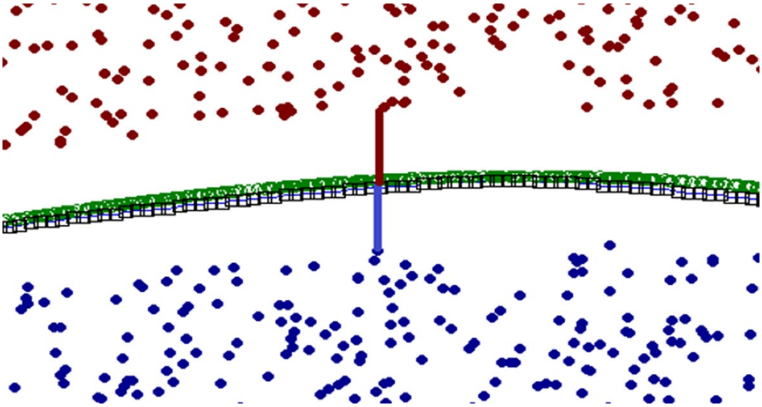


Fig. 6 SVM hypersurface points for a synthetic, linearly separable data set. For illustrative purposes the distance between the points and the hypersurface has been exaggerated. The blue squares and the green circles represent the points identified as belonging to the hypersurface separating the two classes

in Sect. 4, to assess how well the SR model fits the original data. A complementary test can be easily implemented, generating a series of points from the candidate formula and inserting them in the trained SVM. If the equation represents well the boundary between the two classes, the distance of these points from the hypersurface should be close to zero and in particular smaller than the error bars. Indeed, if the points generated by the SR model are in a range of distances from the SVM hypersurface not higher than the uncertainties in the measurements, in practice the obtained equation can be considered a good approximation of the boundary between the two classes. An additional test consists of course of checking that the final model classifies the original data exactly or almost exactly as the SVM.

To better take into account the error bars of the measurements, symbolic regression can be run with the FF including the geodesic distance, according to Eqs. (9) and (10).

7 Numerical tests of SR via GP to obtain boundary equations

The procedure described in the previous section has been subjected to a systematic series of numerical tests. The results have always been positive and the proposed technique has always allowed recovering the original equations describing the boundary between the two classes. In the following Subsection 7.1, an example, of the same level of complexity as the experimental cases presented later, is described in detail. Subsection 7.2 provides some information about the computational requirements to implement the proposed techniques. More details and the main results about these numerical tests are provided in Appendix 1 (Table 2).

7.1 Example including the effect of noise on high dimensional data

As mentioned, there is no conceptual difficulty in applying the proposed methodology to high dimensional problems. Of course, the computational resources required increase exponentially with the number of independent variables (the so called curse of dimensionality). Also the quality of the measurements must be adequate. But these are problems related to the available computational power and/or the quality of the data; in no way they affect the applicability of the proposed technique. Indeed it has been verified with a series of systematic tests that, with adequate level of computer time, high dimensional problems can be solved. As an example, a quite demanding case is reported in the following, for an equation involving 7 variables. The equation used to generate the data is:

Table 2 General GP parameters for the calculation of the boundary equations

GP Parameters	Value(s)
Population size	500
Selection method	Ranking and Tournament
Fitness function	AIC
Constant range	Integers between – 10 and 10
Maximum depth of trees	7
Genetic operators (probability)	Crossover (45%) Mutation (45%) Reproduction (10%)

$$y = x_1 x_2 + \sin(x_3) + \cos(x_4) - x_5/x_6$$

It is worth mentioning that there are many applications of complex science, for example deterministic chaos, in which one has to deal with problems of dimensionality not higher than 7. A total of 4000 points, 2000 per class, have been generated starting from the previous equation. Adding an appropriate level of random noise to the synthetic points (sampled from a Gaussian pdf with standard deviation equal to 10% of the values in the reported examples), forces them to fall on one side of the generating equation, which is meant to simulate a generic boundary. Therefore the generated points become unbiased examples of the two classes. More details about the synthetic data are provided in Table 3. After generating the grid, training the SVM and finding the hyper-surface points, SR via GP has been applied and the following expression for the hyper-surface has been found:

$$y = 0.9\left(x_1 x_2 + \sin(x_3) + \cos(x_4) - x_5/x_6\right) \tag{14}$$

The equation identified by the method is practically the original one. The slightly different multiplicative factor in front is not to be ascribed to a weakness of the method but to the dataset provided as input, since the accuracies of both the SVM and the mathematical equation obtained are equal to 100%. Again, this example proves that, if the surface of the boundary between the cases is sufficiently regular, the dimensionality is not an insurmountable issue, provided enough computational power is available.

The numerical examples presented previously and in Appendix 1include cases where the success rate of the SVM classification is close to 100% (always well in excess of 98% when no differently specified). This is an interesting situation from a scientific point of view; the SVM has learned almost perfectly the boundaries between the classes and therefore the main issue remaining consists of formulating the equations of these boundaries in a mathematical form appropriate for understanding the phenomena. If the data are such that the success rate of classification of the SVM is lower, the proposed method works well anyway, since its objective is the reformulation of the boundary equation found by the SVM. The success rate required for the SVM and the interpretation of the results is an issue, which depends on the application and the objective of the analysis, but does not impact on the validity of the developed technique.

It is worth also restating that, in all the cases tested (see also Appendix 1), even if the final models of the boundaries obtained by the SVM allow classifying with almost 100% accuracy, they have nothing to do with the equations generating the data. Indeed, whatever the actual formula generating the data, the model of the SVM is always of the form of Eq. (1). Therefore in many scientific applications related to the physics of complex systems, whose objective consists of understanding the actual phenomena behind the boundaries and not simply achieving high classification rates, the SVM are not of much use, except when combined with SR via GP, as proposed in this work.

Table 3 The function used to generate the data and the range of variables

Steps:	Values:
Initial function	$y = x_1x_2 + \sin(x_3) + \cos(x_4) - x_5/x_6$
Ranges of variables	$0 < x1 < 2$ & $1.5 < x2 < 3$ $-2 < x3 < 4$ & $0 < x4 < 6$ $4 < x5 < 12$ & $1 < x6 < 4$
Number of nodes for each class	2000

7.2 Computational requirements

To provide an estimate of the computational resources required by the developed methodology, a specific test has been run for the example of 5 variables in Appendix 1. The computer, used to perform the calculations, was an Intel Xeon E5520, with 2 processors, a 2.27 GHz clock, with 8 cores and 24 gigabyte of RAM, with Windows 64 bit operating system. The step of finding the hyper-surface points on a grid of $16^4 * 51$ (16 for the four independent variables and 51 for the dependent one) took 3 h. In its turn, the SR calculation required 48 h, whereas training the SVM took only a few minutes. The SR is therefore by far the task requiring most of the computational resources. On the other hand, it is worth considering that the implemented routines have not been parallelized in any way. Therefore, since both the building of the grid and SR via GP could be easily parallelized, reducing the computational resources even of orders of magnitude is considered a realistic target for future applications.

8 Real world examples

To show the potential of the proposed methodology to attack real life problems, in this section its application to some experimental databases is reported. They have been collected in the framework of various disciplines. The first example is a typical case of a major issue at the frontiers of complexity in Big Physics experiments, namely Magnetic Confinement Nuclear Fusion (MCNF); the determination of the boundary between the safe and disruptive regions of the operational space. For this example, all the various aspects of the proposed method are described, particularised for the case of probabilistic SVM. The other two applications consist of important examples of remote sensing in the field of atmospheric physics and for brevity sake only the main aspects of the technique are covered. The term remote sensing indicates the set of measurement methods aimed at obtaining information about objects without being in contact with them. These techniques can be used to monitor various aspects of the atmosphere and also the effects of human activities on the environment. One example is a case of imagery applied to the assessment of the health of vegetation. The other involves the analysis of laser backscattering signals for the detection of forests fires. For these two cases of application to remote sensing, the traditional SVM method has been implemented. The excellent results obtained in these real life applications prove the value and the flexibility of the proposed methodology.

8.1 The identification of the boundary between disruptive and safe regions of the operational space in Tokamaks

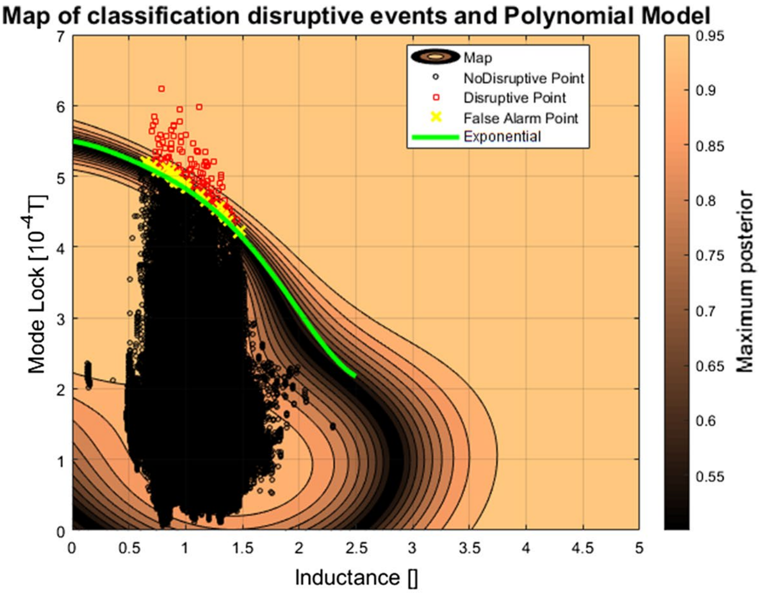
In the last years, collapses and their causes have become not only a major field of research but have also captured the attention of the mainstream media. From market crashes to earthquakes and structural failures in civil engineering, increasing attention is devoted to surprising and typically unexpected abrupt changes in complex systems, leading to catastrophic consequences. The statistical investigation of these phenomena, particularly for robust prediction, requires the development of new mathematical tools (Hadlock 2012). The systematic use of machine learning methods for this purpose is continuously increasing.

Thermonuclear fusion is a field of research aimed at reproducing in the lab the physical process generating energy in the stars, which consists of coalescing light nuclei to generate heavier ones. The very high temperatures required convert the fuel into a specific state of matter called plasma. To confine this ionised gas, the approach using magnetic fields is the most advanced. In Tokamaks, the most successful magnetic configuration to achieve thermonuclear fusion, disruptions remain the most serious cause of collapse. Disruptions are sudden losses of confinement, leading to the abrupt quenching of the plasma with potential major risks for the structural integrity of the devices (Wesson 2004). Since the potential hazard posed by disruptions increases with size, the percent of disruptions allowed in the next generation of devices is quite limited. But disruptions are also a serious issue for the present largest devices. For example, they are one of the main impediments to systematic high current operation in JET (Wesson 2004), (Ongena 2004), (Romanelli 2009), particularly now that the new combination of materials, Be in the main chamber and W in the divertor, renders the first wall less forgiving than in the past.

Given their potential impact on the integrity of the devices, disruptions are a subject of extensive research at present. Various methods of mitigation are being investigated, particularly massive gas injection and shatter pellets (Meitner 2017). The main objective of these techniques consists of limiting the energy conducted directly to the wall by converting the highest percentage of it into radiation. On the other hand, these conversion methods have not only to be effective but are also required not to pose themselves other hazards to the machines, such as excessive increases of the eddy currents due to very fast current quenches. To reduce the strain on the devices also avoidance tactics are being considered, to undertake remedial actions and prevent the occurrence of disruptions. This is particularly important in the perspective of the final reactor, since already in the demonstrative fusion reactor unmitigated disruptions will have to be almost completely avoided and the number of mitigated ones minimised (Wenninger 2016).

Of course, robust prediction tools are a prerequisite to any mitigation or avoidance strategy. Unfortunately, the theoretical understanding of the causes of disruptions is not sufficient to guarantee reliable predictions. As a consequence, existing first principle models are not effective in predicting disruptions on a routine basis. Therefore, in the last decades, a lot of efforts have been devoted to developing empirical models, capable of launching an alarm when a disruption is approaching. Various generations of predictors based on machine learning tools have also been applied to JET data in the last decades. Many alternatives have been explored, ranging from Neural Networks to Self Organizing Maps and fuzzy decision trees (Cannas 2013; Murari 2013; Vega 2009; Gaudio 2014; (Murari 2016, 2017). Unfortunately all these different solutions are practically black boxes, which can help in practice but so far have not contributed much to the understanding of the physics behind disruptions.

To show the potential of the method proposed in this paper to find the boundary between the safe and disruptive regions of the operational space, a large database of JET, including thousands of experiments of the largest device in the world, has been analysed (see Appendix 2). A systematic analysis with the CART approach has shown that, among the global quantities available in real time on JET (including magnetic field, plasma current, safety factor at the edge, plasma beta, diamagnetic energy, radiated power etc.), the locked mode amplitude and internal inductance signals are among the most relevant for disruption prediction. Adding additional quantities practically does not improve the performance of the classifiers. Therefore, also for continuity with the past treatments, they are the two features adopted in this pilot study. The posterior probabilities have then been calculated as indicated in Sect. 3. The adaptive training has been performed for a whole range of threshold



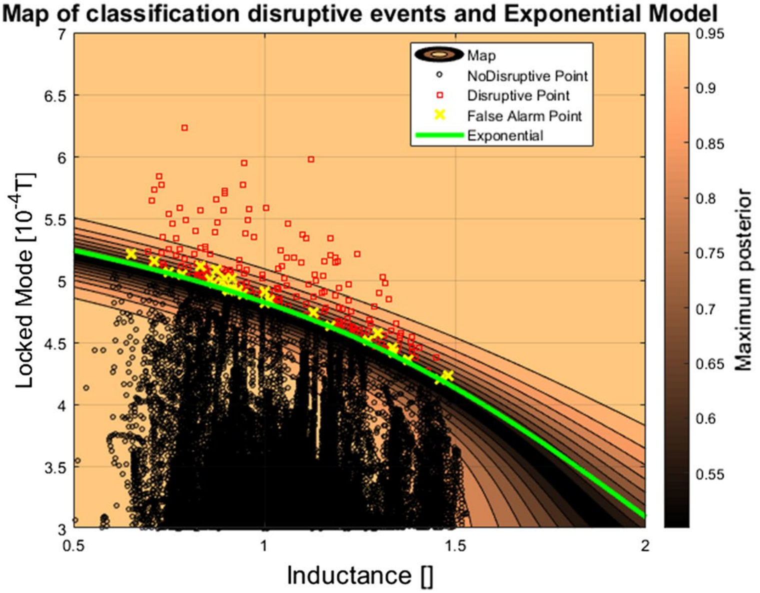


Fig. 7 Top: plot of the safe and disruptive regions of the operational space in JET with the ILW. The colour code represents the posterior probability of the classifier. The black circles are all the non-disruptive shots (10 random time slices for each shot). The red squares are the data of the disruptive shots at the time slice when the predictor triggers the alarm. The yellow crosses are the false alarms. Bottom: zoom of the most relevant boundary region. In thermonuclear fusion, the internal inductance of the plasma is normalised and becomes a dimensionless quantity. Empty brackets on the x axis of the plots indicate this fact

Table 4 The results reported in the row Training refer to the ones obtained by the adaptive training

Model	Success rate	Tardy	Early	Missed	False	Missed + Tardy
TRAINING	96.2% (180/186)	2.7% (5/186)	0.5% (1/186)	0.5% (1/186)	3.9% (40/1016)	3.2% (6/186)
TEST	97.9% (183/187)	2.1% (4/187)	0% (0/187)	0% (0/187)	2.8% (29/1020)	2.1% (4/187)

The ones in the row called Test have been obtained by reapplying the final model obtained at the end of the last campaign back to the entire set of data. The terms Tardy alarms, Missed alarms and Early alarms are defined in Appendix 2

probabilities. It turns out that the probability value, which provides the best performance in terms of success rate, is 60%.

The level plots of the posterior probability obtained are reported in Fig. 7. The curve in light green represents the equation derived with SR via GP (see later). The safe and disruptive regions are well separated in the plane of the locked mode and internal inductance. The clear separation is confirmed by the results in terms of success rate and false alarms reported in Table 4, from which it is easy to appreciate the extremely good performance of the probabilistic SVM. To fully appreciate the results reported in Table 4, it should also be considered that, with the adaptive training from scratch implemented, practicality all the examples inputted to the classifier are to be considered new and never seen before by the SVM [(Murari 2019, (Murari 2020)].

The methodology, described in the Section on Symbolic Regression, has then been applied to the model obtained at the end of the adaptive training. The following model has been retained as a good compromise between complexity and accuracy:

$$y(x) = a_0 \exp\left(a_1 x^{a_2}\right) \tag{15}$$

where y is the locked mode expressed in 10^{-4} Tesla, x the internal inductance and the coefficients assume the values:

$$\begin{aligned} a_0 &= 5.4128 \pm 0.0031; \\ a_1 &= -0.11614 \pm 0.00085; \\ a_2 &= 2.21 \pm 0.011; \end{aligned} \tag{16}$$

The performance of the previous equation, in terms of the usual figures of merit adopted to qualify predictors, reproduces very well the one of the original model as can be appreciated from Table 5.

Comparing Tables 4 with Table 5, it is possible to see how the obtained equation reflects almost exactly the performance of the original model derived by training the probabilistic SVM. In graphical terms, Eq. (15) is shown in light green in Fig. 7; from the plots of this

Table 5 The figures of merit obtained using Eq. (15)

Probability Threshold	Success rate	Tardy	Early	Missed	False
60	97.9% (183/187)	2.1% (4/187)	0% (0/187)	0% (0/187)	2.8% (29/1020)

The terms Tardy alarms, Missed alarms and Early alarms are defined in Appendix 2 figure, it easy to appreciate how the analytical formula, obtained with the proposed methodology, follows almost exactly the 60% curve level of the probabilistic SVM. Therefore, reformulating the equation of the boundary, in a more interpretable way than the output of the SVM, does not imply any significant loss of information in this case. In addition to the good performance, it must be noticed how Eq. (15) represents a major simplification compared to the sum of tens of Gaussians centred on the support vectors, the model of the original SVM. It should also be mentioned that model (15) has also been applied to much more recent campaigns, and therefore to complete new examples, with performance comparable to those of Table 5, confirming the capability of the methodology to provide results of great generalisation potential (Murari [B] 2020). From the point of view of the physics interpretation, Eq. (15) shows how the critical amplitude of the locked mode depends on the internal inductance and therefore on the current profile. In particular, more peaked profile can tolerate a higher level of the locked mode before disrupting. This evidence generalises other treatments, such as the one proposed in (Vries 2015), where it is argued that the amplitude of the locked mode is not the unique quantity to interpret the boundary between the safe and disruptive regions of the operational space.

8.2 Botany: "wilt" database

Healthy vegetation is a prerequisite for the survival not only of many biological communities and ecosystem processes, but of the entire human species. Various remote sensing techniques have become increasingly important for monitoring and understanding the state of the vegetation in many parts of the planet. The reflectance from the foliage in specific ranges of wavelengths allows monitoring the chlorophyll concentration in the vegetation, which is a key factor in determining its health and potential for growth. Healthy vegetation typically absorbs in the red and blue regions of the spectrum, reflects strongly in the near infrared (NIR) and displays strong absorption in wavelengths typical of atmospheric water. Measuring variations in the emission of radiation from vegetation, particularly the ratio between visible and infrared, can provide meaningful information about plant health, environmental stress, and other important characteristics.

For applying our proposed algorithm to real-world health vegetation remote sensing, we selected a database related to botany named "wilt". This database was prepared by Brian Johnson from the Institute of Global Environmental strategies in Japan in 2013 and contains the results of a remote sensing study, for detecting diseased trees, with Qickbird imagery (Johnson 2013). The data set consists of image segments, generated by processing the available pansharpened pictures. The segments contain spectral information from the Quickbird multispectral image bands and texture information from the panchromatic (Pan) image band. In the following, the entries of this database are listed:

Class: 'w' (diseased trees), 'n' (all other land cover).
GLCM_Pan: GLCM mean texture (Pan band).
Mean_G: Mean green value.
Mean_B: Mean blue value.
Mean_NIR: Mean NIR value.
SD_Pan: Standard deviation (Pan band).

This database contains 4339 samples: 74 of them from sick trees and the rest related to other land cover. The new proposed methodology has been applied to this database for finding the classification hyper-surface between the two mentioned classes. The entries have been classified first with the SVM (with the RBF kernel). The subsequent application of our technique to traditional SVM, grid plus SR, has allowed finding the following equation:

$$Mean_B = 22.39 \cdot Mean_G^{0.4705} \tag{17}$$

$$Train\ Accuracy: \ 99.4\% \quad Test\ Accuracy: \ 99.5\ \%$$

The test accuracy value reported has been obtained with the *leave one out* method, chosen because of the very imbalanced character of the database available. Since it presents a success of 99%, practically the same as the SVM, the derived Eq. (17) indicates that the important attributes for classifying this database are the Mean green values and the Mean blue values. Figure 8 reports the entries of the database projected on the plane of these two variables, together with the hyper-surface obtained with Eq. (17).

It is also worth mentioning that, to obtain the same success rate, the SVM has to utilise 1299 support vectors. Therefore the application of the proposed methodology results in a simplification of orders of magnitude in the complexity of the equation, without any significant loss in terms of classification accuracy. Moreover, the obtained formula is susceptible of comparison with models and theoretical considerations, whereas the SVM model is practically intractable from this point of view.

8.3 Remote sensing of the environment: detection of widespread smoke with LIDAR

Light Detection And Ranging (LIDAR) is a remote sensing technique to monitor the atmosphere of a quite long tradition. It was originally proposed at the beginning of the 1960s, shortly after the invention of the laser. LIDAR measuring systems combine laser-focused imaging with the capability of calculating distances, by measuring the time for a signal to return to the point of emission. LIDAR was first developed for meteorology

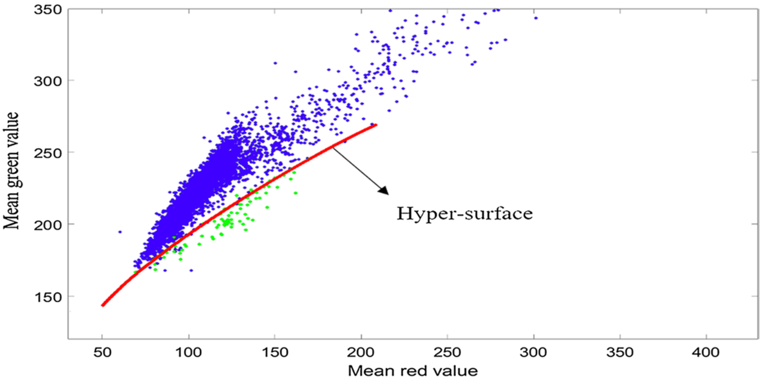


Fig. 8 Distribution of data in the "wilt" database. The green points are diseased trees and the blue points indicate all other types of land cover. The red line indicates the equation obtained for the hyper-surface

but nowadays the approach has become widespread and it is a consolidated technology to make high-resolution maps. The ability to map large areas from a significant distance has allowed the technique to be very useful in many applications from geomatics, archaeology and geography to seismology, forestry and geology. Particularly important have been the results obtained in remote sensing, atmospheric physics and contour mapping.

With regard to remote sensing of the atmosphere, global warming has rendered the problem of wildfires particularly severe in many regions of the world, from the Americas to Australia. In this field, LIDAR has been profitably deployed for the detection of the smoke plume emitted by wild fires. The technique can indeed combine the reliable survey of large areas with the potential for early detection (Fiocco 1963; Andreucci 1993; Bellecci 2007; Bellecci 2010; Vega 2010; Gelfusa 2014). Up to now, the main research efforts have been devoted to improving the accuracy of detecting quite concentrated smoke plumes, which are the main feature exploited for raising alarms during the first stage of wildfires. The operational strategy is based on continuously monitoring the area to be surveyed with a suitable laser; a significant peak in the backscattered signal is considered a sign of a starting fire, justifying the triggering of an alarm. In these circumstance, the strong peaks in the backscattered signal are the features to be detected and various quite successful methods have been refined to this end. In different contexts, it would be useful also to detect widespread smoke, which can be due either to strong winds dispersion or to the presence of non-concentrated sources (Gelfusa 2015). In this application, the measurement requirements are quite different; the signature of the smoke presence is not a strong peak in the detected power but an overall increase over large regions of the signal. Typical backscattered signals for the alternatives of no smoke, strong smoke plume and widespread smoke are reported in Fig. 9.

Starting from the typical Lidar equation (Fiocco 1963), it has been decided to fit the backscattered signal intensity with a mathematical expression of the form:

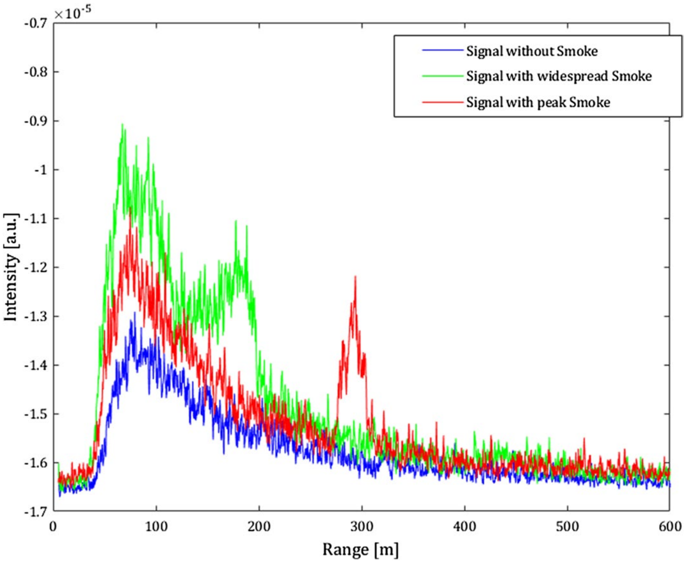


Fig. 9 Examples of LIDAR back scattered signals: **a** Clear atmosphere (blue line) **b** strong smoke plume (red line) **c** widespread smoke (green line)

$$P = \frac{K_1}{R^2} e^{-2K_2 R} \tag{18}$$

where K_1 and K_2 are constants and R is the range. The data of Fig. 9 have been fitted with this formula. The results of the non –linear fit are:

In case of widespread smoke:

$$P = \frac{2.648 \times 10^{-1}}{R^2} \times \exp\left(-1.259 \times 10^{-3} \cdot R\right) \tag{19}$$

Clear atmosphere:

$$P = \frac{1.734 \times 10^{-1}}{R^2} \cdot \exp\left(-1.171 \times 10^{-3} \times R\right) \tag{20}$$

The results of the fit, Eqs. (19) and (20), indicate quite clearly that the parameter K_2 are very similar for both the case of widespread smoke and clear atmosphere. On the other hand, there is a clear difference, of the order of 25% in the constants K_1. This is expected since K_1 includes the effects of the backscattering properties of the atmosphere [(Fiocco 1963), (Bellecci 2010)].

Since the attempt to identify the presence of widespread smoke is a quite pioneering application of the LIDAR technique, it is important not only to be able to discriminate between the two situations but also to provide models for the interpretation of the physics. In particular, the identification of the boundary, in the space of the parameters K_1 and K_2 for the two cases, is considered an essential piece of information for comparison with theories. The proposed methodology has therefore been applied to a quite substantial database:

Total number of data = 521.
number of non-smoke data = 312.
number of widespread smoke data = 209.
number of train data (~ 80%) = 431.
number of test data (~ 20%) = 90.

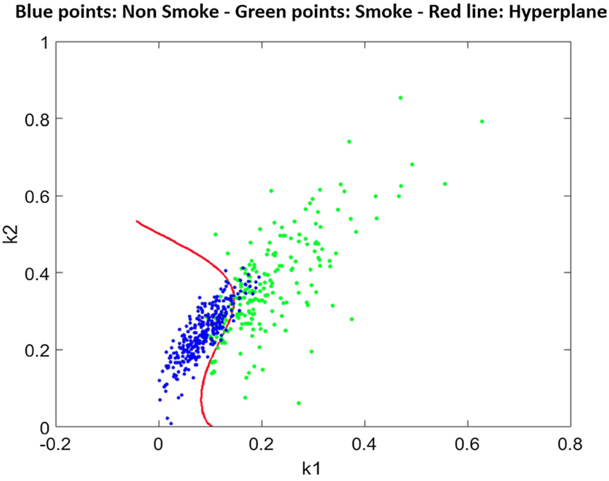


Fig. 10 Eq. (21), describing the boundary between the cases of clear atmosphere and widespread smoke, in the space of the parameters K_1 and K_2

For the SVM, a radial basis functions kernel has been used. The best equation found is:

$$K_1 = 0.1083 \cdot \sin\left(15.61 \cdot K_2^2\right) + 0.1083 \cdot \cos\left(1.5941 \cdot K_2^{0.264}\right) \tag{21}$$

$$\textit{Train Accuracy}: \ 89.33\ \% \qquad \textit{Test Accuracy}: \ 91.11\ \%$$

The equation of the boundary between clear atmosphere and widespread smoke, in the space of the parameters K_1 and $\mathrm{K}_{2,}$ is shown in Fig. 10.

To understand the importance of the results obtained, it should also be considered that the model of the SVM consists of 154 support vectors. Therefore the level of simplification obtained with Eq. (21) is substantial. Moreover, also in this case the formalism of the SVM provides an equation of the boundary between the two classes, which has no relation with the relevant physics.

9 Conclusions

The study of nonlinear complex systems typically requires sophisticated forms of pattern recognition, clustering and in general identification. The arsenal of traditional machine learning tools for classification is very powerful in terms of success rate; on the other hand the models of most available techniques lack physics fidelity and interpretability. In this paper, an original methodology has been presented to obtain the equation of the boundary between two classes, combining almost all the main machine learning techniques available. With the proposed approach, the power of machine learning tools is combined with the realism, physics fidelity and interpretability of equations expressed in the usual formalism of typical scientific theories. In particular, the noise-based ensemble of CART trees has proved essential in identifying the most important features to include in the analysis in an efficient way, taking into account the problem of the noise from the first step of the treatment. The choice of SVM ensures that their structural stability, their capability to maximize the safety margins in the classification, is fully retained in the final results. On the other hand, symbolic regression via genetic programming allows achieving very good physics fidelity and finding a good trade-off between accuracy of classification and complexity of the final equations of the boundary. Therefore the models, obtained with the proposed methodology, are able to better support fundamental scientific activities such as testing of mathematical theories, evaluation of confidence intervals, scaling, extrapolations and experimental design (Murari [B] 2019). It is also worth mentioning that, given the high flexibility of SR via GP, it is normally crucial to exploit all the available “a priori” information about the investigated systems, in order to steer the solutions towards mathematical expressions, which best reflect the actual physical processes of the phenomena under study. The prior knowledge of the scientists can be brought to bear at least on three different stages: on the selection of the most appropriate basis functions, on the definition of the suitable constraints on the structure of the trees and finally on the definition of the fitness function details.

Given the fact that the objectives of the approach are realism and interpretability, a reasonable reduction of the classification performance is not a major issue and can be tolerated. It is also true that symbolic regression via genetic programming can reproduce the accuracy of the classification by the SVM, provided sufficient computational power is available and the data is of adequate quality.

It is also worth emphasizing one more time that the proposed procedure is coherent in the treatment of the error bars of the measurements. This is a very important aspect in the perspective of the application of the developed tools in the domain of complex science, in which small inaccuracies can have dramatic consequences. In the proposed approach, from the noise-based ensemble and the choice of the σ in the Radial Basis Function kernel of the SVM, to the use of the GD as the fitness function for the SR, the effects of the uncertainties in the measurements can be fully taken into account in all the steps of the procedure. For noise of Gaussian distribution, the probabilities obtained have been always satisfactory both in the case of numerical simulations and the analysis of experimental data (Murari [A] 2017, Murari 2012). In this respect, the information theoretic tool of the Geodesic Distance can be quite effective. Of course, the improvements ascribable to GD depends on the individual circumstances but its deployment can significantly improve results, particularly in challenging situations often typical of practical applications, such as presence of high noise level, scarcity of data and significant number of outliers (Murari [B] 2016). For the experimental DBs discussed in Sect. 8, which are of good quality, the degradation consequence of replacing the GD with the RMSE is of the order of a couple of percent points. In any case, a systematic analysis of the improvements, potentially provided by the GD in the case of experimental databases, is beyond the scope of the present paper and is left for future works. Also proving that, in general, the final probabilities are well calibrated, also for noise and perturbations of pdfs different from a Gaussian, remains a serious tasks for further developments.

The numerical tests shown have proved the effectiveness of the proposed technique to identify the real equation of the boundary between classes even in relatively high dimensions, provided the shape of the boundary is a simply connected and sufficiently regular surface. This is very encouraging since, in many scientific applications, the boundaries between the various classes are quite regular functions. This has been confirmed by the application of the methodology to experimental databases of different scientific disciplines, for which sufficient prior information is available to guide the algorithms towards realistic models. On the other hand, in applications to complex system, for example in the identification of attractors, it would be very useful to have a data mining tool adequate to handle more complex and even multiply connected surfaces. It is indeed a topic of future investigations to extend the technique to the investigation of more complex boundaries (Peluso 2014).

The proposed methodology is also susceptible of various additional improvements. First of all, the technique should be extended to other machine learning tools such a neural networks (the only major thread of machine learning not included in the present version of the methodology). Moreover, the task of regression, and not only classification, should also be tackled (Peluso 2014; Murari 2015, 2013). Also applications to various aspects of tomography inversion are envisaged (Marrelli 1998, Martin 1997, Craciunescu 2018).

From a methodological perspective, it should be considered that, even if all the aspects of the proposed procedure are essential, the most delicate is certainly SR via GP, due to the importance of the fitness function in the selection of the best models. Indeed, parsimony pressure in GP tends to focus just on the size of the evolved expression (e.g. (Poli 2003; Luke and Panait 2002)). The model selection criteria implemented do not properly quantify the complexity of the equations, with the consequent danger of overfitting the training data (Raja). For example, (x + x + x) has a larger tree size than sin(x), yet the latter has a far more complex behaviour (Vapnik 2013). Further efforts should therefore certainly be devoted to devising better indicators of the model complexity, to be used in the FF.

Another important theoretical aspect is the extension of the approach to cases affected by noise of different statistics (and not only the usual Gaussian). In this respect, advances in information geometry, with the formulation of geodesic distances valid for other noise statistics, are considered the right direction of future work. From the computational point of view, the heaviest step of the proposed methodology is SR via GP. It is clear that this part of the method is highly parallelizable; therefore much progress is expected, in the reduction of running time, by parallel implementation of SR via GP.

Appendix 1 Numerical Tests for SR via GP to obtain realistic boundary equations

The procedure described in Sect. 6 has been subjected to a systematic series of numerical tests. A significant set of these tests is reported in this Appendix after a detailed description of the numerical method implemented.

The main technique to produce synthetic data and to test the methodology consists of the following 6 steps:

1. Definition of an initial function for the boundary
2. Generating samples of the two classes from the function
3. Training the SVM for classification
4. Building an appropriate mesh on the domain
5. Determining a sufficient number of points on the hyper-surface identified by the SVM
6. Deploying symbolic regression to identify the equation of the hypersurface from the points previously obtained

In the rest of this Appendix, more details about this procedure are provided with the discussion particularised for the case of binary classification and traditional SVM.

In the first step, an initial function as a combination of arithmetic, trigonometric, and exponential operators of independent variables x_i is defined. In general, this function can be written as follows:

$$y = f\left(x_1, x_{2\,\ldots} \right) \quad a_1 < x_1 < b_1 \quad a_2 < x_2 < b_2 \text{etc}$$

In the second step, an adequate number of random points in the valid range of the variables are generated. Then, a positive offset and some random values are added to the y for half of the data to produce the first class; a negative offset and some random values are added to y for the other half to produce the second class. The equations for producing the two classes can be summarized as follow:

$$y_1 = y + \mathit{noise\ of\ standard\ deviations} + \mathit{offset}$$
$$y_2 = y + \mathit{noise\ of\ standard\ deviations} - \mathit{offset}$$

where y_1 and y_2 are the values for the first and second class, respectively.

In the third step, an SVM with "Gaussian Radial Basis Function kernel" is trained. The method used to find the separating hyperplane is "Sequential Minimal Optimization". Depending on the level of random noise, different success rates can be obtained. For the numerical tests presented in the following, the success rate in the classification of the SVM is always very close to 100%.

In the fourth step, a mesh on the domain has to be built in order to identify points sufficiently close to the hypersurface.

The fifth step consists of the identification of the points sufficiently close to the hypersurface, with the algorithm described in Sect. 6.

In the sixth step, the selected hypersurface points are used as inputs to the symbolic regression code, to find the appropriate formula for describing the hypersurface. The settings adopted to run the GP implementing the SR are reported in Table 2.

Examples for two independent variables

Example 1 As a first test, a purely arithmetic function has been tested. The function and ranges of the variables are:

$$y = x_1 + x_2 - x_1 \cdot x_2 \quad \text{where } -1 < x_1 < 1\ \&\ 1 < x_2 < 2$$

After carrying out the six-step procedure described in Sect. 6, the following expression has been obtained:

$$y = 1.011\left(x_1 + x_2 - x_1 \cdot x_2\right)$$

SR via GP converges on a final expression that is in excellent agreement with the initial function, describing the boundary between the two classes. This is particularly true since such a good approximation has been obtained without the non-linear fitting, normally the last step of the SR method.

Example 2 As a second test, a more complex function comprising exponential, arithmetic, and power operators has been assumed for the boundary between the two classes. The function and ranges of the variables are:

$$y = e^{(x_1 \cdot x_2)^{0.5}} \text{ where } 0 < x_1 < 1\ \&\ 1 < x_2 < 3$$

After carrying out the six-step procedure in Sect. 6, the following expression has been obtained:

$$y = 0.974\, e^{(x_1 \cdot x_2)^{0.5}}$$

Again SR via GP converges on a final expression that is in excellent agreement with the initial function describing the boundary between the two classes, even without making recourse to the non-linear fitting step.

Example 3 As the third test, a more complex function comprising trigonometric and arithmetic operators has been defined and 4% classification noise was added to the database. The function and ranges for the variables are:

$$y = \sin\left(x_1\right) + x_2 \text{ where } -3 < x_1 < 3;\ -2 < x_2 < 2$$

After carrying out the six-step procedure in Sect. 6, the following expression has been obtained:

$$y = 0.985\left(\sin\left(x_1\right) + x_2\right)$$

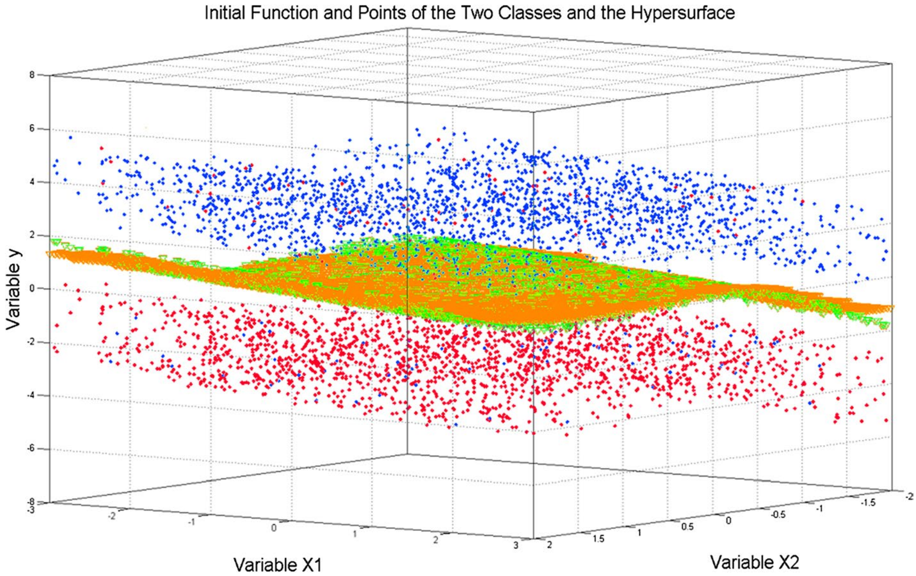


Fig. 11 Points and surfaces of example 3 with two independent variables. The green rectangles are points generated from the initial function, the blue are the points belonging to the first class, the red points are those belonging to the second class, and the orange surface identifies the hyper-surface obtained with the SR via GP

Again SR via GP converges on a final expression that is in excellent agreement with the initial function describing the boundary between the two classes, even without making recourse to the non-linear fitting step. Figure 11 presents the results of this example in pictorial form.

Examples for three independent variables

Some examples considering equations with three independent variables are reported in this section.

Example 1 As a first test, a function comprising only arithmetic operators has been defined. The function and ranges for the variables are:

Initial Defined Function: $y = x_1 - x_2 + x_3$
Range of Variables: $1 < 1 < 2$; $3 < x_2 < 5$; $0 < x_3 < 1$.

The final function obtained from the hypersurface points is:

$$y = 1.002\left(x_1 - x_2 + x_3\right)$$

Example 2 As a second test, a function comprising trigonometric and arithmetic operators has been defined. The function and ranges for the variables are:

Initial Defined Function: $y = x_1 + \sin\left(x_2 \cdot x_3\right)$.
Range of Variables: $1 < x_1 < 2$; $3 < x_2 < 5$; $0 < x_3 < 1$.

The final function obtained from the hypersurface points is:

$$y = 0.98\left(x_1 + \sin\left(x_2 \cdot x_3\right)\right)$$

Again these results confirm the great potential of the approach. Almost exactly the original function can be obtained already at the stage of SR. With additional rounding off of the results or application of non-linear fitting, exactly the original function can easily be recovered.

Example for four independent variables

In this subsection, we describe the results of the application of the SVM-GP methodology to a more complex and noisy database. A five-dimensional synthetic database has been generated with the characteristics described in Table 6.

The procedure for finding the best sigma for the SVM has been applied and the best sigma for the classification is equal to 0.6. The final accuracies of classification for the train and test data are presented in Table 7.

After generating the grid and finding the hyper-surface points, SR via GP has been applied and the following expression for the hyper-surface has been obtained:

$$y = 0.9334\sin\left(0.9190\left(x_1 + x_2\right)\right) - 0.5010\,x_3 \cdot x_4$$

The obtained equation is in good agreement with the initial function. The quality of this estimate can be confirmed by comparing the success rate of the SVM and of the equation

Table 6 Settings for testing SVM-GP on a five-dimensional synthetic database

Steps:	Values:
Initial function	$y = \sin\left(x_1 + x_2\right) - 0.5\,x_3 \cdot x_4$
Ranges of variables	$-1.5 < x_1 < 1.5$ & $-2 < x_2 < 2$ $0 < x_3 < 2$ & $2 < x_4 < 4$
Number of nodes for each class	2000
Thickness of the data's bulk	3
Offset	10% of y domain
Classification noise	~4%

Table 7 The success rates of the SVM for the train and test data on the classification of the synthetic database with the best sigma that equals to 0.6

Database type:	Classification accuracy in percent:
Train data	96.1337
Test data	96.0422

Table 8 The success rates obtained for the train and test data for the classification of the synthetic database with the expression obtained via SR

Database type:	Classification accuracy in percent:
Train data	96.1060
Test data	96.3061

found by SR via GP. The classification success rate of the equation found with SR is reported in Table 8 (to be compared with the results reported in Table 8).

The comparison of the accuracies obtained via SVM and with our proposed technique allows concluding that the SVM-GP approach has excellent performance, even for more complex databases and in higher dimensions, in interpreting the SVM hyper-plane as a hyper-surface equation.

Appendix 2 Database of JET with a metallic wall

All experiments in JET campaigns C29 to C31 have been considered. After proper cleaning and validation of the DB, overall 187 disruptive and 1020 non disruptive shots are included, unless differently specified. JET database with the ILW has been used to implement the methodology described in this paper. In building the database, the intentional disruptions have been excluded from the training. Only time slices, whose plasma current exceeds 750 kA, have been considered but no other general selection has been implemented. All the signals have been resampled at 1kH frequency. Alarms, which are launched 10 ms or less from the beginning of the current quench, are considered tardy, since 10 ms is the minimum time required on JET to undertake mitigation action. Alarms triggered more than 2.5 s before the beginning of the current quench are considered early.

Declarations

Conflict of interest The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

Amari S et al (2000) Methods of information geometry. Translations of mathematical monographs. Oxford University Press

Andreucci F et al (1993) A study on forest fire automatic detection system. Il. Nuovo Cimento 16:35–50. https://doi.org/10.1007/BF02509209

Azad RMA, Ryan C (2014) a simple approach to lifetime learning in genetic programming-based symbolic regression. Evol Comput 22:287–317. https://doi.org/10.1162/EVCO_a_00111

Bahari N. I. S. et al. (2014) Application of support vector machine for classification of multispectral data 2014 IOP Conf. Ser.: Earth Environ. Sci. 20 012038 https://doi.org/10.1088/17551315/20/1/012038

Baseer AZMA (2018) Application of support vector machine models for forecasting solar and wind energy resources: a review. J Clean Prod. https://doi.org/10.1016/j.jclepro.2018.07.164

Beaumont CN et al (2011) Classifying structures in the interstellar medium with support vector machines the g16.05–0.57 supernova remnant. Astrophys J. https://doi.org/10.1088/0004-637X/741/1/14

Bellecci C et al (2007) Application of a CO2 dial system for infrared detection of forest fire and reduction of false alarm. Appl Phys B 87:373–378. https://doi.org/10.1007/s00340-007-2607-9

Bellecci C et al (2010) In-cell measurements of smoke backscattering coefficients using a CO_2 laser system for application to lidar-dial forest fire detection. Opt Eng 49(12):124302. https://doi.org/10.1117/1.3526331

Breiman JFL (1984) Classification and regression trees. Taylor & Francis. https://doi.org/10.1201/9781315139470

Burnham KP et al (2002) Model selection and multi-model inference: a practical information-theoretic approach, 2nd edn. Springer

Cannas B et al (2013) Automatic disruption classification based on manifold learning for real-time applications on JET. Nucl Fusion 53:093023. https://doi.org/10.1088/0029-5515/53/9/093023

Clark JW (2012) Application of support vector machines to global prediction of nuclear properties. Int J Modern Phys B. https://doi.org/10.1142/S0217979206036053

Craciunescu T et al (2018) Maximum likelihood bolometric tomography for the determination of the uncertainties in the radiation emission on JET TOKAMAK. Rev Sci Instrum 89:053504. https://doi.org/10.1063/1.5027880

De Vries PC et al (2014) The influence of an ITER-like wall on disruptions at JET. Phys Plasmas. https://doi.org/10.1063/1.4872017

De Vries PC et al (2015) Scaling of the MHD perturbation amplitude required to trigger a disruption and predictions for ITER. Nucl Fusion 56:026007. https://doi.org/10.1088/0029-5515/56/2/026007

Fiocco G et al (1963) Detection of scattering layers in the upper atmosphere (60–140 km) by optical radar. Nature 199:1275–1276. https://doi.org/10.1038/1991275a0

García S et al (2009) A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. Soft Comput 13:959. https://doi.org/10.1007/s00500-008-0392-y

Gaudio P et al (2014) An alternative approach to the determination of scaling law expressions for the L-H transition in Tokamaks utilizing classification tools instead of regression. Plasma Phys Control Fusion 56:114002. https://doi.org/10.1088/0741-3335/56/11/114002

Gelfusa M et al (2014) UMEL: A new regression tool to identify measurement peaks in LIDAR/DIAL systems for environmental physics applications. Rev Sci Instr 85:063112. https://doi.org/10.1063/1.4883184

Gelfusa M et al (2015) First attempts at measuring widespread smoke with a mobile lidar system. Fotonica AEIT Italian Conference on Photonics Technologies, https://doi.org/10.1049/cp.2015.0187

Hadlock CR (2012) Six sources of Collapse. Mathematical Association of America Washington. https://doi.org/10.4169/j.ctt13x0mx7

Johnson BA et al (2013) A hybrid pansharpening approach and multiscale object-based image analysis for mapping diseased pine and oak trees. Int J Remote Sens 34(20):6969–6982. https://doi.org/10.1080/01431161.2013.810825

Koza JR (1992) Genetic programming: on the programming of computers by means of natural selection. MIT Press, Cambridge

Lungaroni M et al (2018) On the potential of ruled-based machine learning for disruption prediction on JET. Fusion Eng Des 130:62–68. https://doi.org/10.1016/j.fusengdes.2018.02.087

Lungaroni M et al (2019) Geodesic distance on gaussian manifolds to reduce the statistical errors in the investigation of complex systems. Complexity 2019:5986562. https://doi.org/10.1155/2019/5986562

Marrelli L et al (1998) Total radiation losses and emissivity profiles in RFX. Nucl Fusion 38(5):649. https://doi.org/10.1088/0029-5515/38/5/301

Martin P et al (1997) Soft x-ray and bolometric tomography in RFX. Rev Sci Instrum 68(2):1256–1260. https://doi.org/10.1063/1.1147911

Meitner S et al (2017) Design and commissioning of a three-barrel shattered pellet injector for DIII-D Disruption Mitigation Studies. Fusion Sci Technol 72(3):318–323. https://doi.org/10.1080/15361055.2017.1333854

Molnar C (2017) Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/

Murari A et al (2008) Prototype of an adaptive disruption predictor for JET based on fuzzy logic and regression trees. Nucl Fusion. https://doi.org/10.1088/0029-5515/48/3/035010

Murari A et al (2012) A statistical methodology to derive the scaling law for the H-mode power threshold using a large multi-machine database. Nucl Fusion. https://doi.org/10.1088/0029-5515/52/6/063016

Murari A et al (2013) Clustering based on the geodesic distance on Gaussian manifolds for the automatic classification of disruptions. Nucl Fusion. https://doi.org/10.1088/0029-5515/53/3/033006

Murari A et al (2016) A Metric to Improve the Robustness of Conformal Predictors in the Presence of Error Bars. Volume 9653 of the series Lecture Notes in Computer Sciences, pp 105–115. https://doi.org/10.1007/978-3-319-33395-3_8

Murari A et al (2019) A model falsification approach to learning in non-stationary environments for experimental design nature. Sci Rep. https://doi.org/10.1038/s41598-019-54145-7

Murari A et al (2020) (2020) Investigating the physics of Tokamak global stability with interpretable machine learning tools. Appl Sci 10(19):6683. https://doi.org/10.3390/app10196683

Murari A et al (2009) Unbiased and non-supervised learning methods for disruption prediction at JET. Nucl Fusion 49:055028. https://doi.org/10.1088/0029-5515/49/5/055028

Murari A et al (2013) Non-power law scaling for access to the H-mode in tokamaks via symbolic regression. Nucl Fusion 53:043001. https://doi.org/10.1088/0029-5515/53/4/043001

Murari A et al (2015) A new approach to the formulation and validation of scaling expressions for plasma confinement in tokamaks. Nucl Fusion 55:073009. https://doi.org/10.1088/0029-5515/55/7/073009

Murari A et al (2016) Application of transfer entropy to causality detection and synchronization experiments in tokamaks. Nucl Fusion 56:026006. https://doi.org/10.1088/0029-5515/56/2/026006

Murari A et al (2017a) Determining the prediction limits of models and classifiers with applications for disruption prediction in JET. Nucl Fusion 57:016024. https://doi.org/10.1088/0029-5515/57/1/016024

Murari A et al (2017b) Robust scaling laws for energy confinement time, including radiated fraction, in Tokamaks. Nucl Fusion 57:126017. https://doi.org/10.1088/1741-4326/aa7bb4

Murari A et al (2019) Adaptive learning for disruption prediction in non-stationary conditions. Nucl Fusion 59:086037. https://doi.org/10.1088/1741-4326/ab1ecc

Murari A et al (2020) On the transfer of adaptive predictors between different devices for both mitigation and prevention of disruptions. Nucl Fusion 60(5):056003. https://doi.org/10.1088/1741-4326/ab77a6

Ongena J et al (2004) Towards the realization on JET of an integrated H-mode scenario for ITER. Nucl Fusion 44(1):124–133. https://doi.org/10.1088/0029-5515/44/1/015

Peluso E et al (2014) A statistical method for model extraction and model selection applied to the temperature scaling of the L-H transition. Plasma Phys Control Fusion 56:114001. https://doi.org/10.1088/0741-3335/56/11/114001

Platt JC (2000) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola A et al (eds) Advances in large margin classifiers. MIT Press, Cambridge. https://doi.org/10.7551/mitpress/1113.001.0001

Poli R (2003) A simple but theoretically motivated to control bloating in genetic programming" In: Genetic Programming, Proceedings of EuroGP, https://doi.org/10.1007/3-540-36599-0_19

Rattá GA et al (2010) An advanced disruption predictor for JET tested in a simulated real-time environment. Nucl Fusion 50:025005. https://doi.org/10.1088/0029-5515/50/2/025005

Romanelli F et al (2009) Overview of JET results. Nucl Fusion 49(10):104006. https://doi.org/10.1088/0029-5515/49/10/104006

Sahin MÖ et al (2016) Performance and optimization of support vector machines in high-energy physics classification problems. Nuclear Inst Methods Phys Res 838:137–146. https://doi.org/10.1016/j.nima.2016.09.017

Schmid M et al (2009) Distilling free-form natural laws from experimental data. Science 324:81–85. https://doi.org/10.1126/science.1165893

Luke S and Panait L (2002) "Fighting Bloat With Nonparametric Parsimony Pressure" Conference: Proceedings of the 7th International Conference on Parallel Problem Solving from Nature December 2002 https://doi.org/10.1162/EVCO_a_00111

Steinwart I et al (2008) Support Vector Machines. Springer-Verlag, New York. https://doi.org/10.1007/978-0-387-77242-4

Vapnik V (2000) The nature of statistical learning theory. Information Science and Statistics. Springer. https://doi.org/10.1007/978-1-4757-3264-1

Vapnik V (2013) The nature of statistical learning theory. Published by: Springer Science & Business Media, ISBN 1475724403, 9781475724400

Vega J et al (2009) Automated estimation of L/H transition times at JET by combining Bayesian statistics and support vector machines. Nucl Fusion 49(8):085023. https://doi.org/10.1088/0029-5515/49/8/085023

Vega J et al (2010) A universal support vector machines based method for automatic event location in waveforms and video-movies: applications to massive nuclear fusion databases. Rev Sci Instrum 81(2):023505. https://doi.org/10.1063/1.3302629

Vega J et al (2014) Adaptive high learning rate probabilistic disruption predictors from scratch for the next generation of Tokamaks. Nucl Fusion 54:123001. https://doi.org/10.1088/0029-5515/54/12/123001

Vellido A et al (2012) Making machine learning models interpretable. 20th European Symposium on Artificial Neural Networks Bruges, Belgium, April 25-26-27 - ESANN 2012. https://www.i6doc.com/en/book/?GCOI=28001100967420

Wenninger R et al (2016) Power handling and plasma protection aspects that affect the design of the DEMO divertor and first wall. Submitted for publication in Proceedings of 26th IAEA Fusion Energy Conference

Wesson J (2004) Tokamaks. Published by: Clarendon Press Oxford. Third edition. ISBN: 0 19 8509227

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

A. Murari[1] · M. Gelfusa[2] · M. Lungaroni[2] · P. Gaudio[2] · E. Peluso[2]

[1] Consorzio RFX (CNR, ENEA, INFN, Universita' di Padova, Acciaierie Venete SpA), Corso Stati Uniti 4, 35127 Padova, Italy

[2] University of Rome "Tor Vergata", via del Politecnico 1, 00100 Rome, Italy