Classifier-dependent feature selection via greedy methods

Statistics and Computing (2024) 34:151
https://doi.org/10.1007/s11222-024-10460-2

ORIGINAL PAPER

Classifier-dependent feature selection via greedy methods

Fabiana Camattari[1,2] · Sabrina Guastavino[1,2] · Francesco Marchetti[3] · Michele Piana[1,2] · Emma Perracchione[4]

Received: 7 March 2024 / Accepted: 18 June 2024

Abstract
The purpose of this study is to introduce a new approach to feature ranking for classification tasks, called in what follows greedy feature selection. In statistical learning, feature selection is usually realized by means of methods that are independent of the classifier applied to perform the prediction using that reduced number of features. Instead, the greedy feature selection identifies the most important feature at each step and according to the selected classifier. The benefits of such scheme are investigated in terms of model capacity indicators, such as the Vapnik-Chervonenkis dimension or the kernel alignment. This theoretical study proves that the iterative greedy algorithm is able to construct classifiers whose complexity capacity grows at each step. The proposed method is then tested numerically on various datasets and compared to the state-of-the-art techniques. The results show that our iterative scheme is able to truly capture only a few relevant features, and may improve, especially for real and noisy data, the accuracy scores of other techniques. The greedy scheme is also applied to the challenging application of predicting geo-effective manifestations of the active Sun.

Keywords Statistical learning · Machine learning · Classification · Feature selection · Greedy methods

1 Introduction

Greedy algorithms are currently mainly used to iteratively select a reduced and appropriate number of examples according to some error indicators, and hence to produce surrogate and sparse models (Dutta et al. 2021; De Marchi et al. 2005; Santin and Haasdonk 2017; Wenzel et al. 2021, 2023; Wirtz and Haasdonk 2013). The ambition of this paper is to analyze and extend greedy methods to work in the significantly more challenging case of feature reduction, i.e., as the computational core for feature-ranking schemes in the framework of classification issues.

The importance of this application follows from the fact that, as supervised learning models are usually trained on a reduced number of features, the sparsity enhancement is a crucial issue for statistical learning procedures. Most popular feature reduction procedures include Lasso regression (Tibshirani 1996) or variations of the classical Lasso (Group Lasso (Yuan et al. 2006), Adaptive Lasso (Zou 2006), Adaptive Poisson re-weighted Lasso (Guastavino and Benvenuto 2019) to mention a few), linear Support Vector Machine (SVM) feature ranking (Guyon et al. 2002), Fisher score-based schemes (Duda et al. 2012), methods based on mutual information (Peng et al. 2005), Relief and its variants (Robnik-Åikonja and Kononenko 2003). Nevertheless, given

Fabiana Camattari and Sabrina Guastavino have contributed equally to this work.

✉ Sabrina Guastavino
guastavino@dima.unige.it

Fabiana Camattari
camattari@dima.unige.it

Francesco Marchetti
francesco.marchetti@unipd.it

Michele Piana
piana@dima.unige.it

Emma Perracchione
emma.perracchione@polito.it

1 MIDA, Dipartimento di Matematica, Università di Genova, via Dodecaneso 35, 16145 Genova, Italy

2 Osservatorio Astrofisico di Torino, Istituto Nazionale di Astrofisica, via Osservatorio 20, 10025 Pino Torinese, Torino, Italy

3 Dipartimento di Matematica "Tullio Levi-Civita", Università di Padova, via Trieste 63, 35121 Padova, Italy

4 Dipartimento di Scienze Matematiche "Giuseppe Luigi Lagrange" , Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy

Published online: 06 July 2024

any classifier, which can be in principle highly non-linear as e.g. a neural network, none of those algorithms is able to actually capture all the corresponding most relevant features for that classifier. For instance, in the case of Lasso and its generalizations (Freijeiro-González et al. 2022), drawbacks in feature selection ability are shown when there exist non-linear dependence structures.

More in general, all the above mentioned methods identify an *optimal* subset of features based on general patterns in the data. Among them, schemes based on fuzzy information may help in taking account the correlation between features (Yin et al. 2024, 2023). More recently, wrapper and embedded schemes gained popularity; we refer the reader to Bommert et al. (2020) for a general overview. The former use machine learning algorithms to seek for the *optimal* subset of features by considering all possible feature combinations (Bajer et al. 2020), while for the latter, feature selection is integrated or built into the classifier algorithm (Zebari et al. 2020).

In this paper, we propose the so-called greedy scheme that falls in the class of wrapper feature selection methods, but unlike the classical approaches, such as recursive feature elimination (RFE) or recursive feature augmentation (RFA) (Guyon et al. 2002) and forward step-wise selection (James et al. 2023), our method is fully model-dependent and target-based, meaning that any accuracy score can be maximized during the iterative process. Indeed, given any score and any classifier, the feature-based greedy methods iteratively select the most important feature at each step in a classifier-dependent fashion.

At a more theoretical level, this study investigates the effectiveness of the greedy scheme in terms of the Vapnik-Chervonenkis (VC) dimension (Vapnik and Chervonenkis 1971), which is a complexity indicator common to any classifier, such as Feed-forward Neural Networks (FNNs), and it is related to the empirical risk (Bartlett and Mendelson 2002). As a particular instance, we further investigate how greedy methods behave for kernel-based classifiers, such as SVMs (Shawe-Taylor and Cristianini 2004), and in doing so we considered a particular complexity score, known as kernel alignment. These theoretical findings are used on both synthetic and benchmark datasets, showing that with a greedy feature selection, we are able to find a *minimal* set of features without any accuracy loss. Moreover, we apply greedy methods for a case study concerning the classification and prediction of severe geomagnetic events triggered by solar flares.

Solar flares (Piana et al. 2022) are the most explosive manifestations of the active Sun and the main trigger of space weather (Schwenn 2006). They may be followed by coronal mass ejections (CMEs) (Kahler 1992), which, in turn, may generate geomagnetic storms potentially impacting both space and on-earth technological assets (Gonzalez et al. 1994). Data-driven approaches forecasting these events leverage machine learning algorithms trained against historical archives containing physical features extracted from remote-sensing data such as solar images or time series of physical parameters acquired from in-situ instruments (Bobra and Couvidat 2015; Camporeale et al. 2018; Florios et al. 2018; Guastavino et al. 2023; Telloni et al. 2023). These archives systematically provide a huge amount of descriptors and it is currently well-established that this redundancy of information significantly hampers the prediction performances of the classifiers (Campi et al. 2019). Our feature-based greedy scheme is applied in this context, in order to identify among the features the redundant ones and hopefully to improve the classification performances.

The paper is organized as follows. Section 2 introduces our greedy feature selection scheme, which will be motivated thanks to the theoretical analysis in Subsections 2.1 and 2.2. Section 3 describes the application of greedy feature selection to simulated, benchmark and real datasets. Our conclusions are offered in Sect. 4.

2 Greedy feature ranking schemes

Given a set of examples depending on several features, greedy methods are frequently used to find an *optimal* subset of examples and, for such task, since they might be target-dependent, they have already been proved to be effective (see e.g. Wenzel et al. (2021, 2023, 2024)). Here, instead of focusing on the examples, we drive our attention towards the problem of feature selection. To this aim, we considered a binary classification problem with training examples

$$\Xi = X \times Y = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}, \tag{1}$$

where $\boldsymbol{x}_i \in \Omega \subseteq \mathbb{R}^d$ and $y_i \in \mathbb{R}$. For the particular case of the binary classification setting, we fix $y_i \in \{-1, +1\}$.

In the machine learning framework, feature reduction is typically performed by means of linear models, and once the features are identified, non-linear methods like neural networks are applied to predict the given task. However, the fact that some specific features could be useful for some classifier does not imply that the same feature is relevant for any classification model, and this is probably the main weakness of current feature reduction methods in this context. Conversely, our feature-based greedy method (see e.g. Temlyakov (2008) for a general overview) will consist in iteratively selecting the most important feature at each step and in agreement with the considered classifier.

To reach this objective, as usually done, we split the initial dataset $\Xi = X \times Y$ into training and validation sets, respectively denoted by $\mathcal{X} \times \mathcal{Y}$ and $\mathfrak{X} \times \mathfrak{Y}$. Then, at the $k-1$ greedy step $X^{(k-1)}$ will consists of the $k-1$ features that have already been selected (without loosing generalities the

first $k-1$). At the k-th greedy step, on $\mathcal{X}^{(k-1)} \times \mathcal{Y}^{(k-1)}$ we train $d-k$ models $\mathcal{M}_p$ with $x_1, \ldots, x_{k-1}, x_p$, $p = k, \ldots, d$. Then, given an accuracy score μ (the largest the better), we select the k-th feature as

$$x_k = \text{argmax}_{p=k,\ldots,d}\, \mu(\mathcal{M}_p(\mathfrak{X}^{(k-1)}), \mathfrak{Y}^{(k-1)}). \tag{2}$$

We point out that any model can be used in (2), and this implies a totally target-dependent feature selection, which also accounts for the model used to predict a given task.

In the following we investigate the effects of the proposed scheme in terms of VC dimension and for particular instances of kernel learning theory, while a stopping criterion for the algorithm is discussed later in view of the incoming analysis and trade-off remarks.

2.1 The VC dimension in the greedy framework

We consider the dataset (1), where we now suppose that $\Omega = \bigotimes_{k=1}^{d} \Omega^k$ with $\Omega^k = [a_k, b_k] \subset \mathbb{R}$. Given a classifying function $f : \Omega \longrightarrow Y$ we consider the *zero–one loss function*

$$c(\boldsymbol{x}, y, f) = \frac{1}{2}|f(\boldsymbol{x}) - y|,$$

which is 0 if $f(\boldsymbol{x}) = y$ and 1 otherwise. From this loss, we can define the *empirical risk*

$$\hat{e}(\Xi, f) = \frac{1}{n}\sum_{i=1}^{n} c(\boldsymbol{x}_i, y_i, f).$$

Assuming that Ξ is sampled from some fixed unknown probability distribution $p(\boldsymbol{x}, y)$ on $\Omega \times Y$, we note that the empirical risk is the empirical mean value of so-called *generalization risk*, i.e.:

$$e(f) = \int_{\Omega \times Y} c(\boldsymbol{x}, y, f)\, \mathrm{d}p(\boldsymbol{x}, y),$$

i.e., it is the mean value of c averaged over all possible test samples generated by $p(\boldsymbol{x}, y)$, and hence it represents the misclassification probability. However, minimizing the empirical risk does not necessarily correspond to a low generalization risk (refer, e.g., to (Schölkopf and Smola 2002, §5)) or (Vapnik 1998, §5 & §6)). Indeed, this might lead to poor generalization capability in the sense that statistical learning theory already proved that the generalization *capacity* of a given model is somehow inversely related to the empirical risk. Such general idea can be formalized in different ways, such as via the VC dimension. In order to define it, we need to introduce the concept of *shattering*. Let $\Xi_1, \ldots, \Xi_{2^n}$ be all the different datasets obtainable taking all possible configurations of labels assigned to the data. A class $\mathcal{F}$ shatters the set X if for every dataset Ξ_i, $i = 1, \ldots, 2^n$, there exists a function $f : \Omega \longrightarrow Y$, $f \in \mathcal{F}$, such that $\hat{e}(\Xi_i, f) = 0$.

Definition 1 The VC dimension of a class $\mathcal{F}$ of classifying functions is the largest natural number s such that there exists a set X of s examples that can be shattered by $\mathcal{F}$. If such s does not exist, then the VC dimension is ∞.

Let us consider a class $\mathcal{F}$ of classifying functions on Ω whose VC dimension is $s < n$. Then, if $f \in \mathcal{F}$ and $\delta > 0$, the bound

$$e(f) \leq \hat{e}(\Xi, f) + C(s, n, \delta),$$

holds with probability $1 - \delta$, where the so-called capacity term is

$$C(s, n, \delta) = \sqrt{\frac{1}{n}\left(s\left(\log\frac{2n}{s} + 1\right) + \log\frac{4}{\delta}\right)}.$$

The generalization risk (and thus the test error) is bounded by the sum between the empirical risk (that is the training error) and the capacity term of the class, which is monotonically increasing with the VC dimension. If we choose a *poor* class, we get a low VC dimension but possibly a high empirical risk; this situation is usually called *underfitting*. On the other hand, by choosing a *rich* class we can obtain a very small empirical risk, but the VC dimension, and thus the capacity term, is likely to be large; this condition is called *overfitting*. In the following, our purpose is to study how the VC dimension evolves during the greedy steps. It is natural to guess that the capacity of a classifier increases if the information contained in an added feature is considered.

Definition 2 Let $\mathcal{F}$ be a class of binary classifying functions $f : \Omega \longrightarrow Y$. Letting $\boldsymbol{e}_k$ be the k-th cardinal basis vector, we define the k-blind class $\mathcal{F}^{(k)}$, $k \in \{1, \ldots, d\}$, $\mathcal{F}^{(k)} \subseteq \mathcal{F}$ as the class of functions $f^{(k)} : \Omega \longrightarrow Y$ such that

$$f^{(k)}(\boldsymbol{x}) = f^{(k)}(\boldsymbol{x} + \delta\boldsymbol{e}_k),$$

for any $\delta \in \mathbb{R}$ such that $\boldsymbol{x} + \delta\boldsymbol{e}_k \in \Omega$.

For example, consider the class of functions

$$\mathcal{F}_{W,\boldsymbol{b}} := \{f : \Omega \longrightarrow Y \mid f(\boldsymbol{x}) = \tilde{f}(W\boldsymbol{x} + \boldsymbol{b})\},$$

where $\tilde{f}$ is the activation function, W is a $r \times d$ matrix and $\boldsymbol{b}$ is a $r \times 1$ vector, $r \geq 1$. Many well-known classifiers are included in $\mathcal{F}_{W,\boldsymbol{b}}$, such as, neural networks and linear models. In this setting, classifiers in $\mathcal{F}^{(k)}_{W,\boldsymbol{b}}$ can be constructed by restricting to W and $\boldsymbol{b}$ such that $W_{:,k} = \boldsymbol{0}$, where $W_{:,k}$ is the k-th column of W, and $b_k = 0$.

Remark 1 As $\mathcal{F}^{(k)} \subseteq \mathcal{F}$, the fact that that $\text{VC}(\mathcal{F}^{(k)}) \leq \text{VC}(\mathcal{F})$, trivially follows.

In order to formally prove that by adding a feature in the greedy step the obtained classifier cannnot be less expressive (in terms of VC dimension) than the previous one, we introduce two maps:

- $\pi_k : \Omega \longrightarrow \bigotimes_{\substack{i=1 \\ i\neq k}}^{d} \Omega^i$, so that $\pi_k(\boldsymbol{x}) = (x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_d)$, which is a projection.
- $\iota_\alpha : \pi_k(\Omega) \longrightarrow \Omega$, $\alpha \in \Omega^k$, so that $\iota_\alpha(\boldsymbol{x}) = (x_1, \ldots, x_{k-1}, \alpha, x_{k+1}, \ldots, x_d)$, which is injective.

Note that applying $\iota_\alpha \circ \pi_k$ to X has the effect of setting to α the k-th feature for all the examples.

Proposition 1 *X is shattered by $\mathcal{F}^{(k)}$ if and only if $\iota_\alpha(\pi_k(X))$ is shattered by $\mathcal{F}^{(k)}$.*

Proof Any classifier in $\mathcal{F}^{(k)}$ cannot rely on the k-th feature. Precisely, for each $\boldsymbol{x}_i \in X$ we can find $\delta_i \in \mathbb{R}$ so that $\boldsymbol{x}_i + \delta_i \boldsymbol{e} \in \iota_\alpha(\pi_k(X))$. Hence, it is equivalent for any function in $\mathcal{F}^{(k)}$ to shatter X and $\iota_\alpha(\pi_k(X))$. □

For any function $f^{(k)} \in \mathcal{F}^{(k)}$ and $\alpha \in \Omega^k$, we can define a classifier $g : \pi_k(\Omega) \longrightarrow Y$ such that $g(\boldsymbol{x}) = f^{(k)}(\iota_\alpha(\boldsymbol{x}))$. Denoting by $\mathcal{G}$ the class consisting of such functions g, we achieve the following result.

Proposition 2 *$\iota_\alpha(\pi_k(X))$ is shattered by $\mathcal{F}^{(k)}$ if and only if $\pi_k(X)$ is shattered by $\mathcal{G}$.*

Proof Assume that there exists $f^{(k)} \in \mathcal{F}^{(k)}$ that shatters $\iota_\alpha(\pi_k(X))$. Note that the shattering does not rely on the k-th feature, which is constant, and therefore this is equivalent to shatter $\pi_k(\iota_\alpha(\pi_k(X))) = \pi_k(X)$ in a lower-dimensional space by means of a classifier g so that $f^{(k)} = g \circ \pi_k$. Finally, by defining $\boldsymbol{x}^{(k)} = \pi_k(\boldsymbol{x})$, $\boldsymbol{x} \in \iota_\alpha(\pi_k(X))$, we further obtain $\boldsymbol{x} = \iota_\alpha(\boldsymbol{x}^{(k)})$, and therefore $g(\boldsymbol{x}^{(k)}) = f^{(k)}(\iota_\alpha(\boldsymbol{x}^{(k)}))$ for $\boldsymbol{x}^{(k)} \in \pi_k(X)$, which completes the proof. □

Corollary 1 *We have that* $\mathrm{VC}(\mathcal{G}) \leq \mathrm{VC}(\mathcal{F})$.

Proof By putting together Propositions 1 and 2 we can affirm that X is shattered by $\mathcal{F}^{(k)}$ if and only if $\pi_k(X)$ is shattered by $\mathcal{G}$. Note that X and $\pi_k(X)$ have the same cardinality, and therefore $\mathrm{VC}(\mathcal{G}) = \mathrm{VC}(\mathcal{F}^{(k)})$. We conclude the proof by virtue of Remark 1. □

The results in Corollary 1 formalize the idea that by adding a feature in the greedy step the obtained classifier cannot be less expressive than the previous one. Nevertheless, in this greedy context we face a sort of trade-off that deals with the VC dimension: precisely, a high VC-dimension allows the model to fit more complex patterns but may lead to overfitting. Hence, we will discuss later robust stopping criteria for the greedy iterative rule. Now, as a particular case study, we consider SVM classifiers, which are probably the most frequently used ones. Further, being they based on kernels, other capability measures concerning such classifiers can be straightforwardly studied.

2.2 SVM in the greedy framework

Following the SVM literature, we drive our attention towards strictly positive definite kernels $\kappa : \Omega \times \Omega \longrightarrow \mathbb{R}$ that satisfy

$$\int_\Omega \kappa(\boldsymbol{x}, z) v(\boldsymbol{x}) v(z) d\boldsymbol{x} dz \geq 0, \quad \forall v \in L_2(\Omega),$$

for $\boldsymbol{x}, z \in \Omega$. Then, those kernels can be decomposed via the Mercer's Theorem as (see e.g. Theorem 2.2. Fasshauer (2007) p. 107 or Mercer (1909)):

$$\kappa(\boldsymbol{x}, z) = \sum_{k\geq 0} \lambda_k \rho_k(\boldsymbol{x}) \rho_k(z), \quad \boldsymbol{x}, z \in \Omega,$$

where $\{\lambda_k\}_{k\geq 0}$ are the (non-negative) eigenvalues and $\{\rho_k\}_{k\geq 0}$ are the (L_2-orthonormal) eigenfunctions of the operator $T : L_2(\Omega) \longrightarrow L_2(\Omega)$, given by

$$T[v](\boldsymbol{x}) = \int_\Omega \kappa(\boldsymbol{x}, z) v(z) dz.$$

Mercer's theorem provides an easy background for introducing feature maps and spaces. Indeed, for Mercer kernels we can interpret the series representation in terms of an inner product in the so-called *feature space* F, which is a Hilbert space. Indeed, we have that

$$\kappa(\boldsymbol{x}, z) = \langle \Phi(\boldsymbol{x}), \Phi(z) \rangle_F, \quad \boldsymbol{x}, z \in \Omega,$$

where $\Phi : \Omega \longrightarrow F$ is a *feature map*. For a given kernel, the feature map and space are not unique. A possible solution is the one of taking the map $\Phi(\boldsymbol{x}) = \kappa(\cdot, \boldsymbol{x})$, which is linked to the characterization of F as a reproducing kernel Hilbert space; see Fasshauer and McCourt (2015); Shawe-Taylor and Cristianini (2004) for further details. Both in machine learning literature and in approximation theory, radial kernels are truly common. They are kernels for whom there exists a Radial Basis Function (RBF) $\varphi : \mathbb{R}_+ \longrightarrow \mathbb{R}$, where $\mathbb{R}_+ = [0, \infty)$, and (possibly) a shape parameter $\gamma > 0$ such that, for all $\boldsymbol{x}, z \in \Omega$,

$$\kappa(\boldsymbol{x}, z) = \kappa_\gamma(\boldsymbol{x}, z) = \varphi_\gamma(||\boldsymbol{x} - z||_2) = \varphi(r),$$

where $r = ||\boldsymbol{x} - z||_2$. Among all radial kernels, we remark that the Gaussian one is given by

$$\kappa(\boldsymbol{x}, z) = \kappa_\gamma(\boldsymbol{x}, z) = \mathrm{e}^{-\gamma \|\boldsymbol{x} - z\|_2^2} = \mathrm{e}^{-\gamma r^2}.$$

In the following, for simplicity, we omit the dependence on γ, which is also known as scale parameter in machine learning literature.

With radial kernel as well, SVMs can be used for classification purposes and several complexity indicators, such as

the kernel alignment, can be studied in order to have a better understanding of the greedy strategy based on SVM, i.e., when the generic classifier in (2) is an SVM function. The notion of kernel alignment was first introduced by Cristianini et al. (2001) and later investigated in e.g. Wang et al. (2015). Other common complexity indicators related to the alignment can be found in Donini and Aiolli (2017). Given two kernels κ_1 and $\kappa_2 : \Omega \times \Omega \longrightarrow \mathbb{R}$, the empirical alignment evaluates the similarity between the corresponding kernel matrices. It is given by

$$\mathsf{A}(X, \mathsf{K}_1, \mathsf{K}_2) = \frac{(\mathsf{K}_1, \mathsf{K}_2)_\mathsf{F}}{\sqrt{||\mathsf{K}_1||_\mathsf{F}||\mathsf{K}_2||_\mathsf{F}}},$$

where $\mathsf{K}_1 := \mathsf{K}_1(X)$ and $\mathsf{K}_2 := \mathsf{K}_2(X)$ denote the Gram matrices for the kernels κ_1 and κ_2 on X, respectively and

$$(\mathsf{K}_1, \mathsf{K}_2)_\mathsf{F} = \sum_{i,j=1}^{n} \kappa_1(\boldsymbol{x}_i, \boldsymbol{x}_j)\kappa_2(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

The alignment can be seen as a similarity score based on the cosine of the angle. For arbitrary matrices, this score ranges between -1 and 1.

For classification purposes we can define an ideal target matrix as $\mathsf{Y} = \boldsymbol{y}\boldsymbol{y}^\mathsf{T}$, where $\boldsymbol{y} = (y_1, \ldots, y_n)^\mathsf{T}$ is the vector of labels. Then the empirical alignment between the kernel matrix K and the target matrix Y can be written as:

$$\mathsf{A}(X, \mathsf{K}, \mathsf{Y}) = \frac{(\mathsf{K}, \mathsf{Y})_\mathsf{F}}{\sqrt{||\mathsf{K}||_\mathsf{F}||\mathsf{Y}||_\mathsf{F}}} = \frac{(\mathsf{K}, \mathsf{Y})_\mathsf{F}}{n\sqrt{||\mathsf{K}||_\mathsf{F}}}.$$

Such alignment with the target matrix is an indicator of the classification capacity of a classifier. Indeed, to higher alignment scores correspond a separation of the data with a low bound on the generalization error (Wang et al. 2015).

We now prove the following result which will be helpful in understanding our greedy approach.

Theorem 1 *Given two kernels κ_1 and $\kappa_2 : \Omega \times \Omega \longrightarrow \mathbb{R}$, if $||K_2||_F \geq ||K_1||_F$ then $A(X, K_1, Y) \leq A(X, K_2, Y)$.*

Proof By hypothesis we have that:

$$\mathsf{A}(X, \mathsf{K}_1, \mathsf{Y}) = \frac{(\mathsf{K}_1, \mathsf{Y})_\mathsf{F}}{n\sqrt{||\mathsf{K}_1||_\mathsf{F}}} \leq \frac{(\mathsf{K1}, \mathsf{Y})_\mathsf{F}}{n\sqrt{||\mathsf{K}_2||_\mathsf{F}}}.$$

Then, by adding and subtracting $(\mathsf{K}_2, \mathsf{Y})_\mathsf{F}$ at the numerator, and thanks to the linearity of the norm, we obtain:

$$\begin{aligned}\mathsf{A}(X, \mathsf{K}_1, \mathsf{Y}) &\leq \frac{(\mathsf{K1}, \mathsf{Y})_\mathsf{F}}{n\sqrt{||\mathsf{K}_2||_\mathsf{F}}} \\ &= \frac{(\mathsf{K1} - \mathsf{K}_2, \mathsf{Y} - \mathsf{Y})_\mathsf{F}}{n\sqrt{||\mathsf{K}_2||_\mathsf{F}}} + \frac{(\mathsf{K}_2, \mathsf{Y})_\mathsf{F}}{n\sqrt{||\mathsf{K}_2||_\mathsf{F}}} \\ &= \mathsf{A}(X, \mathsf{K}_2, \mathsf{Y}).\end{aligned}$$

□

Considering again Eq. (2), as a corollary of the previous theorem, we have the following result.

Corollary 2 *If κ is a non-increasing radial kernel, then*

$$A(X^{(k)}, K(X^{(k)}), Y) \geq A(X^{(k-1)}, K(X^{(k-1)}), Y).$$

Proof Being $\varphi : \mathbb{R}_+ \longrightarrow \mathbb{R}$ non-increasing, for $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^d$, we obtain

$$\begin{aligned}&\varphi\left(\|\boldsymbol{x} - \boldsymbol{z}\|_2\right) = \\ &= \varphi(\|(x_1, x_2, \ldots, x_k) - (z_1, z_2, \ldots, z_k)\|_2) \leq \\ &\leq \varphi(\|(x_1, x_2, \ldots, x_{k-1}) - (z_1, z_2, \ldots, z_{k-1})\|_2),\end{aligned}$$

which in particular implies that

$$\mathsf{K}_{ij}(X^{(k-1)}) \geq \mathsf{K}_{ij}(X^{(k)}) \geq 0, \quad i, j = 1, \ldots, n.$$

Thus, we get

$$\|\mathsf{K}(X^{(k-1)})\|_\mathsf{F} \geq \|\mathsf{K}(X^{(k)})\|_\mathsf{F},$$

and hence

$$\mathsf{A}(X^{(k)}, \mathsf{K}(X^{(k)}), \mathsf{Y}) \geq \mathsf{A}(X^{(k-1)}, \mathsf{K}(X^{(k-1)}), \mathsf{Y}).$$

□

The result shown in Corollary 2 formalizes again the fact that at each greedy step, the obtained classifier cannot be *less expressive* than the previous one. Note that this kind of feature augmentation strategy via greedy schemes shows some similarities with the so-called Variably Scaled Kernels (VSKs), first introduced in Bozzini et al. (2015) and recently applied in the framework of inverse problems, see e.g. Perracchione et al. (2023, 2021). Indeed, both approaches are based on adding features and both are again characterized by a trade-off between the model capacity, which can be characterized by the kernel alignment, and the model accuracy. To achieve a good trade-off between these two factors we need a stopping criteria for the iterative rule shown in (2).

2.3 Stopping criterion

In actual applications, the greedy iterative algorithm should select, at first, the most relevant features, and then, if no relevant features are available, any accuracy score should saturate. Among several scores μ, a robust one is the so-called True Skill Statistic (TSS) for its characteristic of being insensitive to class imbalance (Bloomfield et al. 2012). Precisely, letting TN, FP, FN, TP respectively the number of true negatives, false positives, false negatives and true positives, the

TSS is defined by:

$$\begin{aligned}\text{TSS}(\text{TN}, \text{FP}, \text{FN}, \text{TP}) = \ & \text{recall}(\text{TN}, \text{FP}, \text{FN}, \text{TP}) \\ & + \text{specificity}(\text{TN}, \text{FP}, \text{FN}, \text{TP}) - 1,\end{aligned}$$

where

$$\text{recall}(\text{TN}, \text{FP}, \text{FN}, \text{TP}) = \frac{\text{TP}}{\text{FN} + \text{TP}}, \tag{3}$$

and

$$\text{specificity}(\text{TN}, \text{FP}, \text{FN}, \text{TP}) = \frac{\text{TN}}{\text{FP} + \text{TN}}. \tag{4}$$

In order to introduce a stopping criterion, we need to point out that we construct a greedy feature ranking by considering, at each step, q splits of the dataset into training and validation sets. Moreover, we now have to denote by $\{x_s\}_{s\in J}$ the $k-1$ features selected at the k-th step of the greedy algorithm, where J is the set of integers associated to the $k-1$ features ($\text{card}(J) = k-1$). Then, at the k-th step of the greedy algorithm, each one of the $d-k$ datasets, composed by the $k-1$ selected features and the added one x_p, for each $p \in I\setminus J$, being $I = \{1, \dots, d\}$, is divided into training and validation sets. We denote such training and validation sets by $\mathcal{X}_{p,h}^{(k-1)} \times \mathcal{Y}_{p,h}^{(k-1)}$ and $\mathfrak{X}_{p,h}^{(k-1)} \times \mathfrak{Y}_{p,h}^{(k-1)}$, for $h = 1, \dots, q$. Hence, once the models $\mathcal{M}_{p,h}$, for each $p \in I \setminus J$ and $h = 1, \dots, q$, have been trained, the k-th feature is chosen so that:

$$x^* = \text{argmax}_{p\in I\setminus J}\, \mu_p^{(k)}, \tag{5}$$

with

$$\mu_p^{(k)} = \frac{1}{q}\sum_{h=1}^{q} \mu(\mathcal{M}_{p,h}(\mathfrak{X}_{p,h}^{(k-1)}), \mathfrak{Y}_{p,h}^{(k-1)}), \tag{6}$$

and where μ is the TSS score. Finally, the new set of features will be given by $\{x_s\}_{s\in J}$, where $J = J \cup \{l\}$, being $x_l = x^*$.

Letting $m^{(k)}$ be the average of the TSS scores computed on different folds at the k-th step and $\sigma^{(k)}$ the associated standard deviation, we stop the greedy iteration at the k-th step if:

$$r^{(k)} = \frac{|m^{(k+1)} - m^{(k)}|}{\sqrt{((\sigma^{(k+1)})^2 + (\sigma^{(k)})^2)}} < \tau, \tag{7}$$

and τ is a given threshold. By doing so, we stop the greedy algorithm when the added feature does not contribute to the accuracy score. In order to better understand this fact, we provide in the following a numerical experiment with synthetic data. Dealing with real data, we might stop the greedy iteration as shown in (7), but then select only the first k^* features, where k^* is

$$k^* = \text{argmax}_{j=1,\dots,k}\, m^{(j)}. \tag{8}$$

We refer the reader to Algorithm 1 for the greedy pseudo-code, while the list of symbols and notations is reported in Table 1.

Table 1 List of notations in the greedy algorithm

Notation	Meaning
$\mathcal{X}_{p,h}^{(k-1)} \times \mathcal{Y}_{p,h}^{(k-1)}$	The h-th training fold, where p is the index of the feature added to $\mathcal{X}_{p,h}^{(k-1)}$, $p \in I \setminus J$.
$\mathfrak{X}_{p,h}^{(k-1)} \times \mathfrak{Y}_{p,h}^{(k-1)}$	The h-th validation fold, where p is the index of the feature added to $\mathfrak{X}_{p,h}^{(k-1)}$, $p \in I\setminus J$.
$\mathcal{M}_{p,k}$	The model trained on $\mathcal{X}_{p,h}^{(k-1)} \times \mathcal{Y}_{p,h}^{(k-1)}$.
μ	A given score (e.g. the TSS).
$r^{(k)}$	The quantity computed at each iteration as in (7) for the stopping criterion.

Inputs: dataset Ξ; tolerance τ; number of folds q; model class $\mathcal{M}$; accuracy score μ.
Outputs: vector of indices of the selected features J.
Initialization: set $k = 1$; $r^{(1)} = \text{Inf}$; $I = \{1, \dots, d\}$; $J = \emptyset$.
while $r^{(k)} \geq \tau$ **do**
  for *each* $p \in I \setminus J$ **do**
    for $h = 1, \dots, q$ **do**
      Train a model $\mathcal{M}_{p,h}$ with $\mathcal{X}_{p,h}^{(k-1)} \times \mathcal{Y}_{p,h}^{(k-1)}$.
      On the validation set compute
      $\mu(\mathcal{M}_{p,h}(\mathfrak{X}_{p,h}^{(k-1)}), \mathfrak{Y}_{p,h}^{(k-1)})$ as in (6).
    end
  Define the new feature $(x_l = x^*)$ as in (5)–(6).
  end
  $k = k + 1$.
  Compute $r^{(k)}$ as in (7).
  Update the set of features $J = J \cup \{l\}$.
end

Algorithm 1: Pseudo-code for the greedy feature ranking algorithm.

3 Numerical experiments

The first numerical experiment aims to numerically show the convergence of the greedy algorithm and the efficacy of the stopping rule. We than test and compare, with state-of-the-art techniques, our scheme on a benchmark dataset. Finally, we will show an application in the context of space weather, which aims to show how this general method is able to deal with real data and infer on the physical aspects of the problem.

3.1 Experiments with a toy dataset

We first focus on the application of the non-linear SVM greedy technique to a balanced simulated dataset constructed as follows: we considered the set $X = \{\boldsymbol{x}_i\}_{i=1}^n$ of $n = 1000$ random points in dimension $d = 15$ sampled from a uniform distribution over $[0, 1)$ and the set of corresponding function values $\{f_{\alpha,i} = f_\alpha(\boldsymbol{x}_i)\}_{i=1}^n$, where $f_\alpha : [0, 1)^d \longrightarrow \mathbb{R}$ is defined as

$$f_\alpha(\boldsymbol{x}) = e^{x_1^2} + e^{x_2} + 3x_3 + 2\cos(x_4 x_5) + 4x_6^2 + 10^\alpha \sum_{j=7}^{d} x_j, \tag{9}$$

and $\alpha \in \{-8, -6, -4, -2\}$. Each $f_{\alpha,i}$ is then labeled according to a threshold value to obtain the set of outputs $Y = \{y_i\}$, i.e., $y_i = 1$ if $f_{\alpha,i}$ is greater than the mean value attained by f_α, and $y_i = -1$ otherwise. From (9) we note that the first 6 features (i.e., x_j for $j = 1, \ldots, 6$) are meaningful for classification purposes when α is lower than -4, while the contribution of the remaining ones is negligible. The classifier used in the following is a SVM model for which both the scale parameter of the Gaussian kernel and the bounding box are optimized via standard cross-validation. The results of using such a classifier into the greedy scheme are reported in Table 2. Such table contains the greedy ranking of the features x_j, $j = 1, \ldots, d$, and the TSS values obtained at each step by averaging over 7 different validation sets. Letting $\tau = 9\text{e} - 2$ be the threshold for the stopping criteria in (7), the greedy algorithm selects the features reported in Table 2, which are above the black solid line. As expected, the algorithm selects only the first six features (the most relevant ones) when α is *small enough* ($\alpha \leq -6$). Then, as soon as the remaining features become more meaningful the greedy selection takes into account more features. In this didactic example we report all the TSS values until the end, to emphasise the robustness of our procedure that correctly identifies the most relevant features.

3.2 Experiments with a benchmark dataset

As a second test we consider a benchmark dataset and we compare different feature extraction algorithms. We take the Breast Cancer Wisconsin (Diagnostic) dataset (Wolberg et al. 1995) free available at the UCI repository at https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic. The classification task consists in predicting whereas a tumor is malignant or benignant basing the considerations on 569 examples made of 30 features which are computed from digitized images of fine needle aspirate of breast masses. Some of the features are, e.g, radius, texture, perimeter and area of the cancer mass.

We then compare our greedy feature selection strategy with LASSO, RFE, and random forest selection, as implemented in the Python scikit-learn package. Again, the classifier used in the following is a SVM model for which both the scale parameter of the Gaussian kernel and the bounding box are optimized via standard cross-validation. The results will show that the greedy strategy, being tailored for the classifier, is able to find really a few relevant features, only 6. Random forest identifies 17 relevant features, while both RFE and LASSO select 24 features, i.e. almost all of them. The accuracy scores returned by all the groups of selected features on 4 test folds are reported in Table 3. Precisely, we compute the TSS as reference score, the Heidke Skill Score (HSS) (Heidke 1926), precision, recall (see Eq. (3)), specificity (see Eq. (4)), F1 score (which is the harmonic mean of precision and recall), and balanced accuracy (which is the arithmetic mean between recall and specificity). We can observe that the SVM classifier trained with only a few greedily selected features is able to achieve about the same accuracy scores than the SVM trained with all, or almost all (RFE, LASSO), features.

With these examples we already proved the ability of the greedy schemes in eliminating the redundant information, and hence in finding a small subset of features. In the next section, we further consider their application to noisy and real data. Moreover, we test the proposed strategy with other classifiers, as neural networks.

3.3 Applications to solar physics: geo-effectiveness prediction

We now focus on a significant space weather application, i.e., the prediction of severe geomagnetic events based on the use of in-situ data. More specifically, data-driven methods addressing this task typically utilize features acquired by in-situ instruments at Lagrangian point L1 (i.e., the Lagrangian point between the Sun and the Earth) to forecast a significant decrease of the SYM-H index, i.e., the expression of the geomagnetic disturbance at Earth (Wanliss and Showalter 2006).

3.3.1 The dataset and the models

The dataset we use consists of a collection of solar wind, geomagnetic and energetic indices. In particular, it is composed by $N = 7888320$ examples and $d = 15$ features sampled at each minute starting from (1-st January 2005) to (31-st December 2019). Below we summarize the features we use:

1. B [nT], the magnetic field intensity, and B_x, B_y and B_z [nT], its three coordinates.
2. V [Km/s], the velocity of the solar wind, and V_x, V_y and V_z [Km/s], its three coordinates.

Table 2 Feature ranking for the greedy scheme on the dataset generated as in (9)

$\alpha = -8$		$\alpha = -6$		$\alpha = -4$		$\alpha = -2$	
x_j	TSS	x_j	TSS	x_j	TSS	x_j	TSS
$\boldsymbol{x_6}$	$\mathbf{0.495 \pm 0.030}$	$\boldsymbol{x_6}$	$\mathbf{0.450 \pm 0.028}$	$\boldsymbol{x_6}$	$\mathbf{0.462 \pm 0.024}$	$\boldsymbol{x_6}$	$\mathbf{0.505 \pm 0.024}$
$\boldsymbol{x_3}$	$\mathbf{0.715 \pm 0.026}$	$\boldsymbol{x_3}$	$\mathbf{0.737 \pm 0.021}$	$\boldsymbol{x_3}$	$\mathbf{0.743 \pm 0.020}$	$\boldsymbol{x_3}$	$\mathbf{0.687 \pm 0.026}$
$\boldsymbol{x_1}$	$\mathbf{0.791 \pm 0.027}$	$\boldsymbol{x_1}$	$\mathbf{0.803 \pm 0.022}$	$\boldsymbol{x_1}$	$\mathbf{0.798 \pm 0.025}$	$\boldsymbol{x_1}$	$\mathbf{0.798 \pm 0.022}$
$\boldsymbol{x_2}$	$\mathbf{0.939 \pm 0.021}$	$\boldsymbol{x_2}$	$\mathbf{0.942 \pm 0.016}$	$\boldsymbol{x_2}$	$\mathbf{0.938 \pm 0.022}$	$\boldsymbol{x_2}$	$\mathbf{0.924 \pm 0.014}$
$\boldsymbol{x_4}$	$\mathbf{0.933 \pm 0.015}$	$\boldsymbol{x_4}$	$\mathbf{0.940 \pm 0.014}$	$\boldsymbol{x_4}$	$\mathbf{0.942 \pm 0.018}$	$\boldsymbol{x_9}$	$\mathbf{0.847 \pm 0.015}$
$\boldsymbol{x_5}$	$\mathbf{0.955 \pm 0.013}$	$\boldsymbol{x_5}$	$\mathbf{0.957 \pm 0.012}$	$\boldsymbol{x_5}$	$\mathbf{0.955 \pm 0.011}$	$\boldsymbol{x_{10}}$	$\mathbf{0.949 \pm 0.014}$
x_{12}	0.952 ± 0.018	x_{12}	0.954 ± 0.011	$\boldsymbol{x_{12}}$	$\mathbf{0.954 \pm 0.009}$	$\boldsymbol{x_5}$	$\mathbf{0.956 \pm 0.013}$
x_{13}	0.951 ± 0.015	x_{13}	0.951 ± 0.013	x_9	0.947 ± 0.013	$\boldsymbol{x_{11}}$	$\mathbf{0.954 \pm 0.014}$
x_9	0.950 ± 0.016	x_9	0.952 ± 0.017	x_{11}	0.941 ± 0.012	$\boldsymbol{x_4}$	$\mathbf{0.951 \pm 0.016}$
x_{14}	0.917 ± 0.016	x_{14}	0.924 ± 0.016	x_8	0.921 ± 0.011	x_{11}	0.931 ± 0.014
x_{10}	0.909 ± 0.018	x_{10}	0.903 ± 0.012	x_{13}	0.903 ± 0.011	x_7	0.905 ± 0.014
x_8	0.904 ± 0.016	x_8	0.904 ± 0.023	x_{14}	0.906 ± 0.014	x_8	0.905 ± 0.016
x_7	0.871 ± 0.015	x_7	0.872 ± 0.012	x_{15}	0.886 ± 0.013	x_{14}	0.862 ± 0.017
x_{11}	0.862 ± 0.024	x_{11}	0.862 ± 0.028	x_7	0.862 ± 0.020	x_{12}	0.859 ± 0.011
x_{15}	0.883 ± 0.015	x_{15}	0.879 ± 0.012	x_{10}	0.873 ± 0.016	x_{15}	0.881 ± 0.017

The selected features are identified by the bold in the table

Table 3 Average scores for the Breast Cancer dataset obtained with SVM using different subsets of features

Metric	All (30 Features)	LASSO (24 Features)	RFE (24 Features)	Random Forest (17 Features)	Greedy (6 Features)
TSS	0.904 ± 0.040	0.905 ± 0.044	0.914 ± 0.034	0.906 ± 0.032	0.922 ± 0.026
HSS	0.916 ± 0.032	0.918 ± 0.039	0.919 ± 0.026	0.908 ± 0.035	0.928 ± 0.016
Precision	0.980 ± 0.012	0.982 ± 0.009	0.961 ± 0.006	0.946 ± 0.030	0.972 ± 0.019
Recall	0.915 ± 0.044	0.915 ± 0.041	0.936 ± 0.037	0.939 ± 0.016	0.939 ± 0.036
Specificity	0.989 ± 0.007	0.990 ± 0.005	0.978 ± 0.004	0.968 ± 0.019	0.983 ± 0.012
F1 score	0.946 ± 0.021	0.947 ± 0.026	0.948 ± 0.017	0.942 ± 0.022	0.954 ± 0.011
Balanced Accuracy	0.952 ± 0.020	0.953 ± 0.022	0.957 ± 0.017	0.953 ± 0.016	0.961 ± 0.013

3. T, the proton temperature, and ρ, the proton density number [cm^{-3}].
4. E_k, E_m, E_t the kinetic, magnetic and total energies.
5. H_m, the magnetic helicity.
6. SYM-H [nT], a geomagnetic activity index that quantifies the level of geomagnetic disturbance.

The first ten features are acquired at the Lagrangian point L1 by in-situ instruments, the energies and the magnetic helicity are adimensional derived quantities, and the SYM-H is measured at Earth. The task considered in what follows consists in identifying the most relevant features used to predict whereas a geomagnetic event occurred, i.e., when the SYM-H is less than −50 nT (label 1), or not (label -1). The dataset at our disposal is highly unbalanced: the rate of positive events is about 2.5%. In order to exploit our data analysis, we first need to fix the notation. We denote by $\tilde{X} = \{\tilde{\boldsymbol{x}}_i\}_{i=1}^N \subseteq \Omega$, where $\Omega \subseteq \mathbb{R}^d$, the set of input samples an by $\tilde{Y} = \{\tilde{y}_i\}_{i=1}^N$, with $\tilde{y}_i \in \{-1, 1\}$, the set of associated labels. The features denoted by $\tilde{x_j}$, $j = 1, \ldots, d$, represent respectively B, B_x, B_y, B_z, V, V_x, V_y, V_z, T, ρ, E_k, E_m, E_t, H_m and the SYM-H.

The analysis is performed with data aggregated by hours, i.e., letting $m = 60$, $n = N/m$ and

$$\boldsymbol{x}_i = \frac{(\sum_{k=i}^{i+m} \tilde{\boldsymbol{x}_k})}{m},$$

we focused on $X = \{\boldsymbol{x}_i\}_{i=1}^n \subseteq \Omega$. Similarly, we define the set of aggregated labels $Y = \{y_i\}_{i=1}^n$.

Given X and Y, the first step of our study consists in using different feature selection approaches to rank the features accordingly to their relevance (see Subsection 3.3.2). After this step, we investigate how these results can be exploited to improve the prediction task (see Subsection 3.3.3). In doing so, we use both SVM and a Feed-forward Neural Network (FNN) in order to predict whether a geo-effective event occurs or not in the next hour. Specifically, the SVM algorithm is trained by performing a randomized and cross-

Table 4 Feature rankings for the greedy schemes on the dataset used for the prediction of geomagnetic solar storms

Greedy ranking (SVM)		Greedy ranking (FNN)	
x_j	TSS	x_j	TSS
SYM-H	**0.703 ± 0.179**	**SYM-H**	**0.936 ± 0.052**
B_z	**0.823 ± 0.121**	**B**	**0.943 ± 0.034**
V	**0.804 ± 0.115**	**E_t**	**0.958 ± 0.039**
E_t	**0.825 ± 0.176**	V_x	0.934 ± 0.078
V_x	**0.853 ± 0.147**		
E_m	0.804 ± 0.184		
B	0.835 ± 0.115		

validated search over the hyper-parameters of the model (the regularization parameter C and the kernel coefficient γ) taken from uniform distributions on $I_C = [0.1, 1000]$ and $I_\gamma = [0.001, 0.1]$ respectively. Instead, the FNN architecture is characterized by 7 hidden layers. The Rectified Linear Unit (ReLU) function is used to activate the hidden layers, the sigmoid activation function is applied to activate the output, and the binary cross-entropy is used as loss function. The model is trained over 200 epochs using the Adam optimizer with learning rate equal to 0.001, with a mini-batch size of 64 examples. In order to prevent overfitting, an L^2 regularization constraint is set as 0.01 in the first two layers. Further, we make use of an early stopping strategy to select the best epoch with respect to the validation loss.

3.3.2 Greedy feature selection approaches

In order to apply efficiently our greedy strategy to both SVM and FNN, we first consider a subset X_p of the original dataset X with a reduced number of examples: we take $p = 3333$ examples. The so-constructed ranking is compared to a state-of-the-art method, i.e., the Lasso feature selection. Precisely, the active set of features returned by Lasso is composed by: B_x, B_y, B_z, V_y, V_z, T, ρ, E_k, E_m, E_t, H_m and the SYM-H. Note that neither V and B, which are physically meaningful for the considered task, are selected by cross-validated Lasso.

In Table 4 we report the results of the greedy feature ranking scheme by using SVM and FNN. In this table, the features are ordered accordingly to the greedy selection. In particular, the greedy iteration stops with all the features reported in the table accordingly to (7), but the selected features are only the ones above the bold line, as in (8). We can note that, the features selected for both SVM and FNN are only a few, and this is due to the fact that greedy schemes are model-dependent and hence are able to truly capture the most significant ones. Interestingly, the features extracted as the most prominent ones are indeed those associated with physical processes involved in the transfer of energy from the CMEs to the Earth's magnetosphere and, thus, with the CME likelihood for inducing geomagnetic storms. B_z, i.e., a southward directed interplanetary magnetic field, is indeed required for magnetic re-connection with the Earth's magnetic field to occur, and thus for the energy carried by the solar wind and/or CMEs to be transferred to the Earth system. In addition, the bulk speed V, or equivalently the radial component of the flow velocity vector V_x, is directly related to the kinetic energy of the solar wind. On the one hand, it is well known that particularly fast particle streams or solar transients can compress the magnetosphere on the sunward side. On the other hand, high levels of magnetic energy (quadratically proportional to the magnetic field intensity) can be converted into thermal energy that heats the Earth's atmosphere, expanding it. In both cases, it appears evident that the transfer of energy, either kinetic or magnetic or total, enabled by the magnetic reconnection between the interplanetary and terrestrial magnetic fields, disrupts the magnetosphere current system, thus causing geomagnetic disturbances. As a conclusion, the extracted features are the physical quantities with the higher expected predictive capability.

We further point out that in order to extract such features, we make use of a validation set and we do not considered any test set. Therefore, the greedy feature extraction is coherently based on the TSS computed on the validation set. Nevertheless, we are now interested in understanding how the selected features work in the prediction (on tests sets) of the original task and with all examples.

3.3.3 Prediction of geomagnetic solar storms events with greedy-selected features

In order to numerically validate our greedy procedure we compare the performances of SVM and FNN trained with respectively: all features, the features returned by Lasso, and the greedily selected features. The comparison is performed by computing several scores (reported in Tables 5 and 6) and by averaging on different splits of the test set. We can observe that for the SVM-based prediction, when using the features extracted with the greedy procedure, we have a remarkable improvement of all accuracy scores. Further, although the performances of the FNN are essentially the same, independently of the feature selection scheme, we note that we are able to achieve the same accuracy scores with only a few features selected ad hoc (3 in this case). This points out again the fact that features extracted by methods, such as Lasso, might be redundant for the considered classifiers. This is even more evident when using the FNN algorithm, which achieves the same accuracy with only 3 greedily selected features. The improvement in terms of accuracy was remarkable only for SVM classifiers, which are known to be less robust then neural networks to *noise*, i.e., redundant information stored in redundant features.

Table 5 Average scores obtained with SVM using different subsets of features

Metric	All (15 features)	LASSO selection (11 features)	Greedy selection (5 features)
TSS	0.679 ± 0.055	0.677 ± 0.088	0.736 ± 0.051
HSS	0.731 ± 0.043	0.739 ± 0.040	0.808 ± 0.021
Precision	0.822 ± 0.117	0.840 ± 0.068	0.909 ± 0.043
Recall	0.683 ± 0.059	0.681 ± 0.090	0.738 ± 0.052
Specificity	0.995 ± 0.005	0.996 ± 0.002	0.998 ± 0.001
F1 score	0.737 ± 0.041	0.745 ± 0.039	0.812 ± 0.021
Balanced accuracy	0.839 ± 0.027	0.839 ± 0.044	0.868 ± 0.026

Table 6 Average scores obtained with FNN using different subsets of features

Metric	All (15 features)	LASSO selection (11 features)	Greedy selection (3 features)
TSS	0.913 ± 0.054	0.917 ± 0.043	0.895 ± 0.054
HSS	0.685 ± 0.105	0.638 ± 0.119	0.669 ± 0.128
Precision	0.577 ± 0.153	0.519 ± 0.159	0.571 ± 0.176
Recall	0.935 ± 0.065	0.945 ± 0.056	0.919 ± 0.068
Specificity	0.978 ± 0.014	0.972 ± 0.017	0.976 ± 0.019
F1 score	0.695 ± 0.010	0.650 ± 0.114	0.680 ± 0.122
Balanced accuracy	0.957 ± 0.027	0.959 ± 0.022	0.948 ± 0.027

4 Conclusions and future work

We introduced a novel class of feature reduction schemes, namely greedy feature selection algorithms. Their main advantage consists in the fact that they are able to identify the most relevant features for any given classifier. We studied their behavior both analytically and numerically. Analytically, we could conclude that the models constructed in such a way cannot be less expressive than the standard ones (in terms of VC dimension or kernel alignment). Numerically, we showed their efficacy on a problem associated to the prediction of geomagnetic solar storms. As the activity of the Sun is cyclic, work in progress consists in using greedy schemes to study which features are relevant on either high or low activity periods. Finally, as there is a growing interest in physics-informed neural networks (PINN), we should investigate, both theoretically and numerically, which are the challenges that greedy methods could achieve in this context.

Acknowledgements Fabiana Camattari and Emma Perracchione kindly acknowledge the support of the Fondazione Compagnia di San Paolo within the framework of the Artificial Intelligence Call for Proposals, AIxtreme project (ID Rol: 71708). Sabrina Guastavino was supported by the Programma Operativo Nazionale (PON) "Ricerca e Innovazione" 2014–2020. The research by Michele Piana was supported in part by the MIUR Excellence Department Project awarded to Dipartimento di Matematica, Università di Genova, CUP D33C23001110001. Emma Perracchione acknowledges the support of the project NODES within the MUR - M4C2 1.5 of PNRR, grant agreement no. ECS00000036. All authors are members of the Gruppo Nazionale per il Calcolo Scientifico - Istituto Nazionale di Alta Matematica (GNCS - INdAM).

Author Contributions F.C., S.G. and E.P. worked at the implementation of the computational strategy. F.M. and E.P. contributed to the mathematical formulation of the method. S.G., F.C., M.P. and E.P. worked at the manuscripts drafting. S.G., F.C. and M.P. contributed to the formulation and the design of the experiments. All authors collaborated to conceiving the general scientific ideas at the basis of the study.

Funding Open access funding provided by Università degli Studi di Genova within the CRUI-CARE Agreement.

Data availibility The combined solar wind and geomagnetic data analyzed in this paper are public and can be freely downloaded from the NASA's Space Physics Data Facility (http://omniweb.gsfc.nasa.gov/).

Declarations

Conflict of interest The authors declare no competing interests.

References

Bajer, D., Dudjak, M., Zorić, B.: Wrapper-based feature selection: how important is the wrapped classifier? In: 2020 International Conference on Smart Systems and Technologies (SST), pp. 97–105 (2020). IEEE

Bartlett, P.L., Mendelson, S.: Rademacher and gaussian complexities: risk bounds and structural results. J. Mach. Learn. Res. **3**, 463–482 (2002)

Bloomfield, D.S., Higgins, P.A., McAteer, R.T.J., Gallagher, P.T.: Toward reliable benchmarking of solar flare forecasting methods. The Astrophys. J. Letters **747**(2), 41 (2012)

Bobra, M.G., Couvidat, S.: Solar flare prediction using sdo/hmi vector magnetic field data with a machine-learning algorithm. Astrophys J. **798**(2), 135 (2015)

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., Lang, M.: Benchmark for filter methods for feature selection in high-dimensional classification data. Comput. Stat. & Data Anal. **143**, 106839 (2020)

Bozzini, M., Lenarduzzi, L., Rossini, M., Schaback, R.: Interpolation with variably scaled kernels. IMA J. Numer. Anal. **35**, 199–219 (2015)

Campi, C., Benvenuto, F., Massone, A.M., Bloomfield, D.S., Georgoulis, M.K., Piana, M.: Feature ranking of active region source properties in solar flare forecasting and the uncompromised stochasticity of flare occurrence. Astrophys J **883**(2), 150 (2019)

Camporeale, E., Wing, S., Johnson, J.: Machine Learning Techniques for Space Weather. Elsevier, United States (2018)

Cristianini, N., Shawe-Taylor, J., Elisseeff, A., Kandola, J.: On kernel-target alignment. In: Dietterich, T., Becker, S., Ghahramani, Z. (eds.) Advances in Neural Information Processing Systems, vol. 14. MIT Press, Cambridge (2001)

De Marchi, S., Schaback, R., Wendland, H.: Near-optimal data-independent point locations for radial basis function interpolation. Adv. Comput. Math. **23**, 317–330 (2005)

Donini, M., Aiolli, F.: Learning deep kernels in the space of dot product polynomials. Machine Learn. **106**, 1245–1269 (2017)

Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley-interscience, New York (2012)

Dutta, S., Farthing, M.W., Perracchione, E., Savant, G., Putti, M.: A greedy non-intrusive reduced order model for shallow water equations. J. Comput. Phys. **439**, 110378 (2021)

Fasshauer, G.E.: Meshfree Approximations Methods with MATLAB. World scientific, Singapore (2007)

Fasshauer, G.E., McCourt, M.: Kernel-based Approximation Methods Using MATLAB. World scientific, Singapore (2015)

Florios, K., Kontogiannis, I., Park, S.-H., Guerra, J.A., Benvenuto, F., Bloomfield, D.S., Georgoulis, M.K.: Forecasting solar flares using magnetogram-based predictors and machine learning. Sol. Phys. **293**(2), 28 (2018)

Freijeiro-González, L., Febrero-Bande, M., González-Manteiga, W.: A critical review of lasso and its derivatives for variable selection under dependence among covariates. Internat. Stat. Rev. **90**(1), 118–145 (2022)

Gonzalez, W., Joselyn, J.-A., Kamide, Y., Kroehl, H.W., Rostoker, G., Tsurutani, B.T., Vasyliunas, V.: What is a geomagnetic storm? J. Geophys. Res. Space Phys. **99**(A4), 5771–5792 (1994)

Guastavino, S., Benvenuto, F.: A consistent and numerically efficient variable selection method for sparse Poisson regression with applications to learning and signal recovery. Stat. Comput. **29**(3), 501–516 (2019)

Guastavino, S., Candiani, V., Bemporad, A., Marchetti, F., Benvenuto, F., Massone, A.M., Mancuso, S., Susino, R., Telloni, D., Fineschi, S., Piana, M.: Physics-driven machine learning for the prediction of coronal mass ejections' travel times. The Astrophys. J. **954**(2), 151 (2023)

Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learn. **46**, 389–422 (2002)

Heidke, P.: Berechnung des erfolges und der gute der windstarkevorhersagen im sturmwarnungsdienst. Geogr. Ann. **8**, 301–349 (1926)

James, G., Witten, D., Hastie, T., Tibshirani, R., Taylor, J.: An Introduction to Statistical Learning with Applications in Python, pp. 233–235. Springer, Cham (2023)

Kahler, S.: Solar flares and coronal mass ejections. Ann. Rev. Astron. Astrophys. **30**(1), 113–141 (1992)

Mercer, J.: Functions of positive and negative type and their connection with the theory of integral equations. Phil. Trans. Royal Society **209**, 415–446 (1909)

Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. (2005). https://doi.org/10.1109/TPAMI.2005.159

Perracchione, E., Massone, A.M., Piana, M.: Feature augmentation for the inversion of the Fourier transform with limited data. Inverse Probl. **37**(10), 105001 (2021)

Perracchione, E., Camattari, F., Volpara, A., Massa, P., Massone, A.M., Piana, M.: Unbiased CLEAN for STIX in Solar Orbiter. The Astrophys. J. Suppl. Series **268**(2), 68 (2023)

Piana, M., Emslie, A.G., Massone, A.M., Dennis, B.R.: Hard X-ray Imaging of Solar Flares, vol. 164. Springer, Berlin (2022)

Robnik-Šikonja, M., Kononenko, I.: Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learn **53**(1–2), 23–69 (2003)

Santin, G., Haasdonk, B.: Convergence rate of the data-independent P-greedy algorithm in kernel-based approximation. Dolomites Res. Notes Approx. **10**(2), 68–78 (2017)

Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA (2002)

Schwenn, R.: Space weather: the solar perspective. Living Rev. Sol. Phys. **3**(1), 1–72 (2006)

Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge (2004)

Telloni, D., Lo Schiavo, M., Magli, E., Fineschi, S., Guastavino, S., Nicolini, G., Susino, R., Giordano, S., Amadori, F., Candiani, V., et al.: Prediction capability of geomagnetic events from solar wind data using neural networks. The Astrophys. J. **952**(2), 111 (2023)

Temlyakov, V.N.: Greedy approximation. Acta Numer **17**, 235–409 (2008)

Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol. **50**(1), 267–288 (1996)

Vapnik, V.N.: Statistical Learning Theory. Wiley, NY, USA (1998)

Vapnik, V.N., Chervonenkis, A.Y.: On the uniform convergence of relative frequencies of events to their probabilities. Theory Probab. Appl. **16**, 264–280 (1971)

Wang, T., Dongyan, Z., Shengfeng, T.: An overview of kernel alignment and its applications. Artificial Intell. Rev. **43**(2), 179–192 (2015)

Wanliss, J.A., Showalter, K.M.: High-resolution global storm index: Dst versus sym-h. J. Geophys. Res. Space Phys. (2006). https://doi.org/10.1029/2005JA011034

Wenzel, T., Santin, G., Haasdonk, B.: A novel class of stabilized greedy kernel approximation algorithms: convergence, stability and uniform point distribution. J. Approx. Theory **262**, 105508 (2021)

Wenzel, T., Santin, G., Haasdonk, B.: Analysis of target data-dependent greedy kernel algorithms: convergence rates for f-, f·P- and f/P-greedy. Constructive Approx. **57**(1), 45–74 (2023)

Wenzel, T., Marchetti, F., Perracchione, E.: Data-driven kernel designs for optimized greedy schemes: a machine learning perspective. SIAM J. Sci. Comput. **46**(1), 101–126 (2024)

Wirtz, D., Haasdonk, B.: A vectorial kernel orthogonal greedy algorithm. Dolomites Res. Notes Approx. **6**, 83–100 (2013)

Wolberg, W., Mangasarian, O., Street, N., Street, W.: Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository (1995)

Yin, T., Chen, H., Yuan, Z., Wan, J., Liu, K., Horng, S.-J., Li, T.: A robust multilabel feature selection approach based on graph structure considering fuzzy dependency and feature interaction. IEEE Trans. Fuzzy Sys. **31**(12), 4516–4528 (2023)

Yin, T., Chen, H., Wan, J., Zhang, P., Horng, S.-J., Li, T.: Exploiting feature multi-correlations for multilabel feature selection in robust multi-neighborhood fuzzy β covering space. Inf. Fusion **104**, 102150 (2024)

Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Series: B Stat. Meth. **68**(1), 49–67 (2006)

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., Saeed, J.: A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. J. Appl. Sci. Tech. Trends **1**(2), 56–70 (2020)

Zou, H.: The adaptive lasso and its oracle properties. J. American Stat. Association **101**(476), 1418–1429 (2006)