Interpretation of Linear Classifiers by Means of Feature Relevance Bounds

Christina Göpfert, Lukas Pfannschmidt, Jan Philip Göpfert, Barbara Hammer

Cognitive Interaction Technology
Inspiration 1
33619 Bielefeld
Germany

Abstract

Research on feature relevance and feature selection problems goes back several decades, but the importance of these areas continues to grow as more and more data becomes available, and machine learning methods are used to gain insight and interpret, rather than solely to solve classification or regression problems. Despite the fact that feature relevance is often discussed, it is frequently poorly defined, and the feature selection problems studied are subtly different. Furthermore, the problem of finding all features relevant for a classification problem has only recently started to gain traction, despite its importance for interpretability and integrating expert knowledge. In this paper, we attempt to unify commonly used concepts and to give an overview of the main questions and results. We formalize two interpretations of the all-relevant problem and propose a polynomial method to approximate one of them for the important hypothesis class of linear classifiers, which also enables a distinction between strongly and weakly relevant features.

Keywords: Feature Relevance, Feature Selection, Interpretability, All-Relevant, Linear Classification

1. Introduction

Feature relevance and feature selection have been active research areas for many years [1, 2]. However, the impact of these fields only continues to grow as data becomes more and more abundant, and insight into and interpretation of models and frameworks are regarded as more and more important [3, 4, 5], in particular in the light of easily fooled machine learning models [6]. Despite

Email addresses: `cgoepfert@techfak.uni-bielefeld.de` (Christina Göpfert), `lpfannschmidt@techfak.uni-bielefeld.de` (Lukas Pfannschmidt), `jgoepfert@techfak.uni-bielefeld.de` (Jan Philip Göpfert), `bhammer@techfak.uni-bielefeld.de` (Barbara Hammer)

the fact that feature relevance is often discussed in the literature [2, 7], it is frequently poorly defined, and there are subtle differences between the feature selection problems studied in various papers. In addition, the problem of identifying all features relevant to a classification problem has only recently started to gain traction, despite its importance for interpretability and integrating expert knowledge.

Early concepts of feature relevance were developed e.g. by Gennari, Langley and Fisher [8] and Kohavi and John [1]. The definitions by Kohavi and John continue to be used to this day, and form the basis of our analysis. Regarding feature selection, one branch of research is motivated by the fact that the presence of many irrelevant or correlated features can severely impact the speed and generalization ability of a machine learning algorithm. The identification of feature subsets that allow for good classification performance was the subject of the 2003 NIPS feature selection challenge [9]. A wide array of filter, wrapper and embedded methods to solve this problem have been proposed, including Lasso, Group Lasso or Cluster Elastic Net for regression and l_1- or l_1 and l_2-regularized SVM for classification, filters based on mutual information for nonlinear models, or techniques based on relevance learning of variables [1, 10, 11, 12, 13, 14, 15, 16].

More recently, the problem of finding *all* relevant features has become a point of interest, motivated by a desire to use machine learning not only as a blind toolbox for classification or regression, but to understand in detail the behavior of a machine learning model, to integrate expert knowledge, or even to use machine learning in order to explore dependencies within the data. Unlike popular methods such as lasso, which identify only one minimal set of relevant features, the all-relevant feature-selection problem aims for an identification of all features which can be relevant for a given learning task; this is of particular interest in the case of feature correlations and redundancies where researchers might be interested in subtle markers which are otherwise shadowed by the more pronounced signals. The identification of all relevant features enables an interactive expert evaluation to decide which one of a set of highly correlated features is most reasonable in a given setting.

Methods that have been proposed for tackling the all-relevant feature-selection problem include Boruta [17, 18], which uses random forests to calculate importance measures for each feature, forward-backward selection schemes using various relevance measures, or, recently, the calculation of relevance intervals for linear regression and metric learning [19, 20]. To some extent, Group Lasso and Elastic Net are also capable of giving a relevance ranking in the case of mutually redundant features in regression problems [14]. By relying on random forests as a universal approximator, Boruta addresses the problem of identifying all relevant features for the given classification task as a general problem. In contrast, Elastic Net and the relevance learning approach as proposed in the work [19, 20] focus on feature relevance for linear regression or classification, respectively, disregarding possible nonlinear dependencies of features and output variable. Since linear models constitute a particularly relevant model class, this restriction of feature relevances constitutes an important specialization of the general problem. Interestingly, the Elastic Net can be accompanied by mathematical guarantees under which model selection consistency holds [21]. In contrast, the approach for feature relevance in metric learning by Schulz et al. [20], which deals with classification rather than regression, regards the valid interpretation of a specific given model only.

In this paper, we propose a novel method to identify all relevant features for the hypothesis class of linear classifiers, and we derive a polynomial time learning algorithm for this task. More specifically, we address the more general problem of identifying all possible relevances of a given feature for any model with a given shape (e.g. linear) and small error for a given classification problem. The proposed method produces *relevance intervals* that indicate, in the case of linear models, the different levels of importance a feature is assigned by some linear classifier with low error. The benefit of these relevance intervals is that they not only offer a way to determine all relevant features, but they also enable a clear distinction between strongly and weakly relevant features for the given linear classification problem, a distinction that is typically missing in raw relevance profiles. We rely on two approximations: First, we formalize the objective as a constrained optimization problem which controls the classification error on the given data as well as the model's generalization ability by limiting a norm of the weights, as is common in computational learning theory for linear systems. Secondly, we quantify the observed feature relevance by the used feature weight, which is also a common practice for linear models. Based on these two approximations, a mathematical formalization of the problem of determining feature relevance bounds becomes possible.

The remainder of this paper is organized as follows: Section 2 gives an introduction into the concept of feature relevance and formalizes the two main feature selection problems: the *minimal-optimal* and the *general all-relevant* problems. We introduce the new concepts of the *specific all-relevant* problem as well as strong and weak relevance to a hypothesis class. In Section 3 we present a novel method for solving the specific all-relevant problem in the case of linear classifiers, by relying on two steps: First, an initial linear classifier is determined, namely an l_1-SVM, which enables us to find bounds for the quality which can be reached in the given setting. Secondly, for each feature, a minimization and maximization, respectively, of the feature relevance is computed over all linear models with a similar quality as the initial one. We phrase these latter problems as constrained optimization problems, and we show that they can be rephrased as linear problems, i.e. the solution can be found in polynomial time. Section 4 contains experiments on artificial data where we demonstrate the behavior of the model and its superiority to alternatives such as Boruta or Elastic Net for the linear case. Further, we evaluate the stability of the model as compared to initial SVM solutions on real-world data.

2. Feature Relevance and Feature Selection Problems

In this section we give a short introduction to the existing theory of feature relevance and the types of feature selection problems typically encountered in the literature. We extend the existing theory by introducing Definitions 5 and 6 that explore relevance for hypothesis classes.

2.1. Feature Relevance Theory

First, we introduce the notation used in the remainder of this paper. The starting point of our analyses is a binary classification data set

$$\{(x^1, y^1), \dots, (x^n, y^n)\} \subset \mathbb{R}^d \times \{-1, 1\}$$

made up of data vectors x^i and corresponding labels y^i. The (x^i, y^i) are assumed to be independent observations of the random variables (X, Y), $X = (X_1, \dots, X_d)$, with distribution $\mathcal{D}$ over $\mathbb{R}^d \times \{-1, 1\}$. A machine learning algorithm is defined by an *inducer* $\mathcal{I}$ that maps a training sample to some *classification rule* or *hypothesis* $h : \mathbb{R}^d \to \{-1, 1\}$ whereby the set $\mathrm{Im}(\mathcal{I})$ of classification rules the inducer can map to is called the *hypothesis space* $\mathcal{H}$ of $\mathcal{I}$. An inducer typically attempts to find a classification rule that minimizes the *generalization error*

$$L_{\mathcal{D}}(h) = \mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) \neq y] = \mathcal{D}(\{(x, y) : h(x) \neq y\}).$$

We call the $X_1, \dots, X_d$ the *features* of the classification problem and the j-th entry x_j of a data point x the *value of feature j for x*.

The study of the relevance of features to a classification problem can be motivated by improving the prediction performance of the predictors, making predictors quicker and cheaper or gaining a better understanding of the underlying processes of data generation and model functionality [2]. Due to these diverse motivations and the difficulty in rigorously defining relevance, the current literature deals with a broad spectrum of interpretations of feature relevance.

Firstly, it is necessary to distinguish between two areas of possible relevance, namely:

1. The relevance of a feature to the label variable Y, or
2. the relevance of a feature to the behavior of a particular classification rule.

Concerning the relevance of a feature to the label variable Y, in the following we use the definitions given by Kohavi and John [1] where S_j denotes the set of all features except X_j, i.e.

$$S_j = \{X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_d\},$$

and for $S = \{X_{i_1}, \dots, X_{i_{|S|}}\} \subseteq \{X_1, \dots, X_d\}$ and $\boldsymbol{s} \in \mathbb{R}^{|S|}$, $S = \boldsymbol{s}$ denotes the event $X_{i_j} = s_j$ for $j = 1, \dots, |S|$.

Definition 1. A feature X_j is *strongly relevant* to Y if there exists some $x_j \in \mathbb{R}$, $y \in \{-1, 1\}$ and $\boldsymbol{s_j} \in \mathbb{R}^{d-1}$ for which $\mathbb{P}(X_j = x_j, S_j = \boldsymbol{s_j}) > 0$ such that

$$\mathbb{P}(Y = y | X_j = x_j, S_j = \boldsymbol{s_j}) \neq \mathbb{P}(Y = y | S_j = \boldsymbol{s_j}).$$

It is *weakly relevant* to Y if it is not strongly relevant, but can be made strongly relevant by removing other features, i.e. there exists a subset of features S' of S_j for which there exists some x_j, y and $\boldsymbol{s'}$ with $\mathbb{P}(X_j = x_j, S' = \boldsymbol{s'}) > 0$ such that

$$\mathbb{P}(Y = y|X_j = x_j, S' = \boldsymbol{s'}) \neq \mathbb{P}(Y = y|S' = \boldsymbol{s'}).$$

A feature is *relevant* if it is either strongly or weakly relevant. Otherwise, it is *irrelevant*.

The distinction between strong and weak relevance is inspired by the observation that some features may carry information on the predictor variable that is made redundant by the information contained in other features. As an extreme case, consider a dataset where some features are identical copies of one another, such as a dataset with features (X_1, X_2, X_2). Assume that the data can be accurately classified by calculating $X_1 + X_2$. Even though each feature contains information relevant to the classification problem, calling one of the identical copies relevant would be misleading, as would calling one of them irrelevant. In the framework created by Definition 1, the second and third features are weakly relevant, indicating their redundancy, while the first is strongly relevant.

The relevance of a feature to the behavior of a particular hypothesis h is given by Nilsson et al. [22]:

Definition 2. A feature X_j is relevant to the hypothesis h if

$$\mathbb{P}(h(X_j, S_j) \neq h(X'_j, S_j)) > 0$$

where X_j and X'_j are independent samples from the marginal distribution of the feature X_j.

That is, a feature is considered relevant to a particular hypothesis if resampling the feature according to its marginal distribution affects the behavior of the classifier with a non-zero probability. Note that unlike Definition 1, Definition 2 does not distinguish between strong and weak relevance. Such a distinction would require an additional degree of freedom. In Definition 6, we will extend the relevance framework of Kohavi and John [1] to hypothesis classes, and in particular introduce the concepts of strong and weak relevance to a hypothesis class.

2.2. Feature Selection Problems

So far, we have been interested in individual features and assessing their relevance either to a target variable or to a hypothesis. Now, we turn our attention to feature selection problems, where we investigate subsets of features and attempt to choose subsets that fulfill some criteria. There are two types of feature selection problems typically referred to in the literature: the *minimal-optimal* feature selection problem and the *all-relevant* feature selection problem. While the minimal-optimal problem is related to improving the performance of an algorithm [2], the all-relevant problem aims at insight into the data generation and classification processes [17]. Unfortunately, the all-relevant problem is also computationally intractable [22].

Minimal-Optimal

The goal of the minimal optimal problem is usually to improve prediction performance, or to make predictors quicker and cheaper. It is typically considered in terms of a fixed machine learning algorithm and its associated hypothesis class. It can be formalized as follows:

Definition 3. The *minimal-optimal feature selection problem* for the inducer $\mathcal{I}$ is the problem of finding a small subset

$$S \subset \{X_1, \ldots, X_d\}$$

of features such that applying $\mathcal{I}$ to the data set restricted to S incurs a hypothesis h with small generalization error $L_{\mathcal{D}}(h)$ with high probability.

Note that some authors define the minimal-optimal problem as the problem of finding a minimal-size feature subset $S \subseteq \{X_1, \ldots, X_d\}$ such that $\mathbb{P}(Y|S) \approx \mathbb{P}(Y|X_1, \ldots, X_d)$. The result is a smaller feature set on which the optimal Bayes classifier shows identical or similar performance to the original problem. However, this formulation does not take into account biases and trade-offs particular to the inducer $\mathcal{I}$, and thus may not be optimal for improving the performance of the hypotheses learned by $\mathcal{I}$. For example, if the hypothesis class of $\mathcal{I}$ is the set of linear classifiers, a feature subset S on which the optimal Bayes classifier performs well can still lead to failure of the inducer $\mathcal{I}$ if data restricted to S is no longer linearly separable.

In the earlier example of the dataset with features (X_1, X_2, X_2) and optimal classification through calculation of $X_1 + X_2$, a minimal-optimal set consists of either the first and second or the first and third feature. This immediately shows that a minimal-optimal set is not necessarily unique.

An intuitive approach to solve the minimal-optimal problem in an embedded manner for linear classifiers is to apply l_0-regularization. Since this is usually computationally intractable, l_1-regularization is used as an approximation, such as in the Lasso [11] and Elastic Net [14] methods. Nilsson et al. [22] propose a backward-elimination wrapper approach that, for strictly positive data distributions, identifies the minimal-optimal set for the optimal Bayes classifier in the large-sample limit in polynomial time. Ideally, an algorithm that solves the minimal-optimal problem finds all features relevant to the best hypothesis in $\mathcal{H}$ in the sense of Definition 2, and only a subset of weakly relevant features.

All-Relevant

In contrast to the minimal-optimal problem, the all-relevant problem is usually motivated by the need to identify features that are “significant” to the target variable [17], either in order to further investigate their dependencies, e.g. to find exploratory directions in gene micro-array research [23, 24], or in order to enable a more interactive model design process, e.g. to design classifiers that take into account expert knowledge and the costs of acquiring each feature.

When the goal is to further investigate dependencies between the features and the target variable, there is no formal reason to take into account a specific hypothesis class or inducer, since the intended results are independent of potential machine learning applications on the data. Indeed, Definition 1 is sufficient to define the feature set that researchers who perform this type of analysis aim to find:

Definition 4. The *general all-relevant problem* is the problem of identifying all features relevant to the target variable Y in the sense of Definition 1, that is, all strongly and weakly relevant features.

In the literature, this problem is frequently referred to as simply the *all-relevant problem* [17, 22].

While solving the general all-relevant problem is a suitable approach to gain insight into the underlying data distribution and identify possible directions for further research, it is not appropriate when the objective is to facilitate interactive model design or analysis: features that are relevant to the target variable cannot be leveraged by all types of models and thus are not always relevant for model design. Furthermore, even a feature that is irrelevant to the target variable may improve the performance of some models by effectively enlarging the hypothesis class (although this may be an undesirable effect). An example of this was given by Kohavi and John as the hypothesis class of linear classifiers without offset (homogeneous halfspaces), which can effectively be enlarged to include linear classifiers with offset (inhomogeneous halfspaces) by adding an additional feature that takes a constant non-zero value. For these reasons, we define a new all-relevant feature selection problem, taking into account a fixed hypothesis class and leveraging the concept of relevance to a hypothesis introduced in Definition 2.

Definition 5. The *specific all-relevant problem* for a hypothesis class $\mathcal{H}$ is the problem of determining all features relevant in the sense of Definition 2 to some hypothesis $h \in \mathcal{H}$ such that the generalization error $L_{\mathcal{D}}(h)$ is small. More formally, if we fix $\varepsilon > 0$ and define

$$\mathcal{H}_\varepsilon := \{h \mid L_{\mathcal{D}}(h) \leq \varepsilon\},$$

the set of all hypotheses in $\mathcal{H}$ with generalization error at most ε, then the specific all-relevant problem is the problem of identifying all features X_j such that there exists $h \in \mathcal{H}_\varepsilon$ with X_j relevant to h in the sense of Definition 2.

The specific all-relevant problem has, to the best of our knowledge, not been formally considered. However, methods such as Boruta [18, 25] that aim at approximating a solution to the general all-relevant problem can also be interpreted as an attempt to solve the specific all-relevant problem for the hypothesis class they employ. On the other hand, it may be worthwhile to use a solution of the specific all-relevant problem as an approximation of the general all-relevant problem when the latter proves too difficult.

Inspired by Definition 2, we propose the following taxonomy of *relevance to a hypothesis class*:

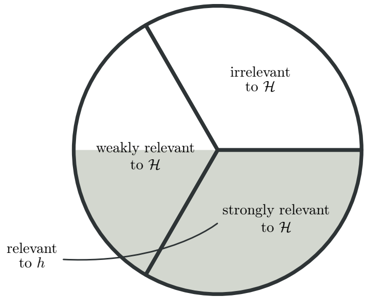

Figure 1: The relationship between the features strongly relevant, weakly relevant, and irrelevant to a hypothesis class $\mathcal{H}$ and the features relevant to some $h \in \mathcal{H}$ (gray areas).

Definition 6. A feature is called strongly relevant to a hypothesis class $\mathcal{H}$ if it is relevant to all $h \in \mathcal{H}$ in the sense of Definition 2. It is weakly relevant to a hypothesis class $\mathcal{H}$ if it is relevant to at least one $h \in \mathcal{H}$, but not all, and relevant if it is either strongly or weakly relevant. It is irrelevant if it is not relevant to any $h \in \mathcal{H}$.

The relationship between the features strongly and weakly relevant to $\mathcal{H}$ and the features relevant to a single $h \in \mathcal{H}$ is visualized in Figure 1. Using Definition 6, we can reformulate the specific all-relevant problem for the hypothesis class $\mathcal{H}$ analogously to Definition 4:

The specific all-relevant problem for the hypothesis class $\mathcal{H}$ is the problem of finding all features that are relevant to $\mathcal{H}_\epsilon$ in the sense of Definition 6.

The distinction between strongly and weakly relevant features is necessary for application domains such as classifier design taking into account expert knowledge and feature acquisition costs: Any hypothesis $h \in \mathcal{H}$ with low generalization error $L_{\mathcal{D}}(h) \leq \varepsilon$ must use all features that are strongly relevant to $\mathcal{H}_\varepsilon$, but only some that are weakly relevant to $\mathcal{H}_\varepsilon$. Thus, knowledge of the set of strongly and weakly relevant features for $\mathcal{H}_\varepsilon$ immediately provides insight into which trade-offs are possible, which are not, and which features cannot be leveraged by the hypothesis class at all. However, we emphasize once more that even though the concept of strong and weak relevance to a hypothesis class is inspired by the concept of strong and weak relevance to a target variable, one does not imply the other, and relevance to the target variable does not even imply relevance to the optimal Bayes classifier, as argued by Nilsson et al. [22]: if the optimal Bayes classifier predicts the same label for all points, it acts independently of all features, even though some features may be relevant to the target variable.

In the following, we propose an approach to solve the specific all-relevant problem for the hypothesis class of linear classifiers, which constitute a popular model choice in the biomedical domain [23].

3. All-Relevant Determination Using Relevance Bounds

In the remainder of this paper, we introduce a novel approach to solve the specific all-relevant problem for the hypothesis class of linear classifiers, by the name of FeReL (**Fe**ature **Re**levance for **L**inear Classification). Our method calculates *relevance bounds* for each feature, which admit the discrimination of strongly and weakly relevant features for the hypothesis class as well as the identification of irrelevant features. We further show that our proposed relevance bounds can be calculated by solving linear programs, and thus our method runs in polynomial time and the results are unique.

3.1. Relevance Bound Intuition

Since we do not have access to the underlying data distribution, we must estimate two quantities: Firstly, whether or not a feature is relevant to a particular hypothesis, and secondly, which features in the hypothesis class induce a low generalization error. Shortly, for the relevance of a feature to a hypothesis, i.e. to a hyperplane defined by a normal vector and an offset, we use as a quantitative measure the absolute values of the normal vector entries, and for the generalization error of a hypothesis, we use a proxy based on the l_1-norm of the normal vector as well as margin intrusions. These ideas will be described in more detail in the remainder of this section.

The heuristics we use to compensate for the fact that the underlying data distribution is unknown are the following: If a linear classifier, that is, a hyperplane, is defined by the normal vector $\boldsymbol{w}$ and offset b, we take the absolute value $|w_j|$ as a measure for the relevance of the feature X_j. In particular, X_j is relevant to the hypothesis $(\boldsymbol{w}, b)$ iff $|w_j| > 0$.[1] Based on this interpretation of relevance, we define *relevance intervals* for each feature in the following way:

Definition 7. The *relevance interval* for the feature X_j is defined as

$$\left[\min_{(\boldsymbol{w},b)\in\mathcal{H}_\varepsilon} |w_j|, \max_{(\boldsymbol{w},b)\in\mathcal{H}_\varepsilon} |w_j| \right].$$

Going back to Definition 6, a feature X_j with relevance interval $[w_{lower}, w_{upper}]$ is strongly relevant to $\mathcal{H}_\varepsilon$ if $w_{lower} > 0$, irrelevant if $w_{upper} = 0$ and weakly relevant if $w_{lower} = 0$ and $w_{upper} > 0$.

Determining the hypothesis class $\mathcal{H}_\varepsilon = \{h \mid L_{\mathcal{D}}(h) < \varepsilon\}$ is complicated by the fact that we cannot exactly determine the generalization error of any hypothesis in our class. Furthermore, the smallest generalization error achievable by our hypothesis class is unknown, so it is unclear how ε should be chosen. Here, we propose the following approach:

[1] This is a common practice [23], and coincides with Definition 2 in many practical cases, e.g. if the features are subject to independent and unbounded noise.

1. Compute a baseline hypothesis h^* using an established machine learning algorithm, e.g. a Support Vector Machine.
2. Let ε be an upper bound for the generalization error of h^*, as given e.g. through Rademacher complexities. Then, as a proxy for $\mathcal{H}_\varepsilon$, use the set $\hat{\mathcal{H}}_\varepsilon$ of hypotheses with the same or a similar upper bound for the generalization error.

Note that our general approach is not specific to linear classifiers and can be extended to any hypothesis class for which risk bounds can be efficiently controlled and an accepted measure for the relevance of a feature to a given classifier exists.

Since one application of interest is designing classifiers that use few, cheap features, we want to encourage sparse weight vectors, and allow importance to "shift" between features in order to gain full information about groups that can be substituted for each other. To this end, we use an l_1-regularized SVM as a baseline linear classifier and set $\hat{\mathcal{H}}_\varepsilon$ to the set of hyperplanes $(\boldsymbol{w}, b)$ with similar hinge loss and l_1-norm $\|\boldsymbol{w}\|_1$. By controlling these two quantities, Rademacher complexities give risk bounds similar to the bounds for the original l_1-regularized SVM solution.

3.2. A Formal Relevance Bounds Method

Concretely, our method consists of the following steps: Given data

$$(x^1, y^1), \dots, (x^n, y^n) \in \mathbb{R}^d \times \{-1, 1\},$$

1. A baseline linear classifier is given by a solution to the l_1-regularized SVM optimization problem:

$$\begin{aligned} \left(\tilde{\boldsymbol{w}}, \tilde{b}, \tilde{\boldsymbol{\xi}}\right) \in \underset{\boldsymbol{w}, b, \boldsymbol{\xi}}{\arg\min} \ & \|\boldsymbol{w}\|_1 + C \sum_{i=1}^{n} \xi_i \\ \text{s. t. } & y_i(\boldsymbol{w}^\top x_i - b) \geq 1 - \xi_i \\ & \xi_i \geq 0, \ i = 1, \dots, n. \end{aligned}$$

   From this baseline classifier, we derive an upper bound on the generalization error we allow that depends on the l_1-norm μ and hinge loss ρ of the baseline classifier. Specifically, we set

$$\mu = \|\tilde{\boldsymbol{w}}\|_1 \text{ and } \rho = \sum_{i=1}^{n} \tilde{\xi}_i.$$

2. As a proxy for $\mathcal{H}_\varepsilon$, we use

$$\hat{\mathcal{H}}_\varepsilon := \{(\boldsymbol{w}, b) \mid \|\boldsymbol{w}\|_1 \leq (1 + \delta) \cdot \mu \text{ and hinge loss } \leq \rho\}.$$

   We will demonstrate in Section 3.3 that this allows us to control an upper bound on the generalization error of the hypotheses in $\hat{\mathcal{H}}_\varepsilon$ – i.e., ε itself

– as a function of δ and and upper bound on the performance of the baseline classifier. We enforce "$\leq$" constraints instead of "$=$" constraints for greater stability and because hyperplanes with a smaller hinge loss and l_1-norm admit the same upper bound for the generalization error. For each feature i, the *minimum feature relevance bound* is then defined as the optimal value of the optimization problem

$$\begin{aligned} \mathrm{minRel}((x_i,y_i)_{i=1}^n, j) : \min_{\boldsymbol{w},b,\boldsymbol{\xi}} \; & |w_j| \\ \text{s. t. } & y_i(\boldsymbol{w}^\top x_i - b) \geq 1 - \xi_i, \xi_i \geq 0,\; i = 1,\dots,n \\ & \sum_{i=1}^n \xi_i \leq \rho,\; \|\boldsymbol{w}\|_1 \leq (1+\delta)\cdot\mu. \end{aligned}$$

The *maximum feature relevance bound* is defined as the optimal value of the optimization problem

$$\begin{aligned} \mathrm{maxRel}((x_i,y_i)_{i=1}^n, j) : \max_{\boldsymbol{w},b,\boldsymbol{\xi}} \; & |w_j| \\ \text{s. t. } & y_i(\boldsymbol{w}^\top x_i - b) \geq 1 - \xi_i, \xi_i \geq 0,\; i = 1,\dots,n \\ & \sum_{i=1}^n \xi_i \leq \rho,\; \|\boldsymbol{w}\|_1 \leq (1+\delta)\cdot\mu. \end{aligned}$$

3. A feature is relevant to the hypothesis class of linear classifiers if its maximum feature relevance bound is greater than zero and irrelevant otherwise. A relevant feature is strongly relevant to the hypothesis class of linear classifiers if the minimum feature relevance bound is also greater than zero, and weakly relevant to the hypothesis class of linear classifiers if the minimum feature relevance bound is equal to zero.

The maximum feature relevance bound of a feature X_j is greater than zero if and only if there exists a hypothesis $h \in \hat{\mathcal{H}}_\varepsilon$ parameterized by normal vector $\boldsymbol{w}$ and offset b such that $|w_j| > 0$. The minimum feature relevance bound is greater than zero if and only if this holds for every $h \in \hat{\mathcal{H}}_\varepsilon$. This means that if $|w_j|$ is a good measure of relevance and $\hat{\mathcal{H}}_\varepsilon$ is a good approximation of $\mathcal{H}_\varepsilon$, our method solves the all-relevant problem for $\mathcal{H}_\varepsilon$, the class of hyperplanes that separate data from the distribution $\mathcal{D}$ as well as our baseline l_1-SVM solution. We justify our choice of $\hat{\mathcal{H}}_\varepsilon$ using generalization bounds based on Rademacher averages in the following section.

3.3. Generalization Bounds

We stated in subsection 3.1 that our minRel and maxRel consider the separating hyperplanes where Rademacher complexities give similar risk bounds as for the output of the l_1-regularized SVM. To see that this is indeed the case, recall Theorem 26.15 of Understanding Machine Learning [26]:

Theorem 1. *Suppose that $\mathcal{D}$ is a distribution on $X \times Y$ such that with probability 1 we have $\|x\|_\infty \leq R$. Let $\mathcal{H} = \{\boldsymbol{w} \in \mathbb{R}^d \mid \|\boldsymbol{w}\|_1 \leq B\}$ and let $l : \mathcal{H} \times X \times Y$ be of the form $l(\boldsymbol{w}, (x, y)) = \varphi(\langle \boldsymbol{w}, x\rangle, y)$ where $\varphi : \mathbb{R} \times Y \to \mathbb{R}$ is such that for all $y \in Y$, the scalar function $a \mapsto \varphi(a, y)$ is η-Lipshitz and such that $\max_{a \in [-B\cdot R, B\cdot R]} |\varphi(a, y)| \leq c$. Then, for any $\tau \in (0, 1)$ with probability of at least $1 - \tau$ over the choice of an i.i.d. sample of size n, for all $\boldsymbol{w} \in \mathcal{H}$,*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[l(\boldsymbol{w}, x, y)] \leq \frac{1}{n}\sum_{i=1}^{n} l(\boldsymbol{w}, x_i, y_i) + 2\eta B R\sqrt{\frac{2\log(2d)}{n}} + c\sqrt{\frac{2\ln(2/\tau)}{n}}$$

We will use Theorem 1 to show that with high probability, the generalization error of every $h \in \hat{\mathcal{H}}_\varepsilon$ can be bounded similarly to the generalization error of the baseline l_1-SVM solution.

Consider the ramp loss

$$l(\boldsymbol{w}, x, y) = \min\{1, \max\{0, 1 - y(\boldsymbol{w}^\top x)\}\}.$$

The ramp loss is 1-Lipshitz and maps to the interval $[0, 1]$. It upper bounds the 0-1 loss, so that using Theorem 1 gives

$$L_{\mathcal{D}}(\boldsymbol{w}) \leq \frac{1}{n}\sum_{i=1}^{n} l(\boldsymbol{w}, x_i, y_i) + 2BR\sqrt{\frac{2\log(2d)}{n}} + \sqrt{\frac{2\ln(2/\tau)}{n}} \tag{1}$$

for all $\boldsymbol{w}$ such that $\|\boldsymbol{w}\|_1 \leq B$ with probability $1 - \tau$ over the choice of sample. In particular, setting $\rho = \sum_{i=1}^{n} \tilde{\xi}_i$ to the hinge loss of the baseline classifier and using the fact that the hinge loss upper bounds the ramp loss, (1) gives the bound

$$L_{\mathcal{D}}(\tilde{\boldsymbol{w}}, \tilde{b}) \leq \frac{\rho}{n} + 2\|\tilde{\boldsymbol{w}}\|_1 R\sqrt{\frac{2\log(2d)}{n}} + \sqrt{\frac{2\ln(2/\tau)}{n}}$$

for the generalization error of the baseline linear classifier $(\tilde{\boldsymbol{w}}, \tilde{b})$ and

$$L_{\mathcal{D}}(h) \leq \frac{\rho}{n} + 2(1+\delta)\|\tilde{\boldsymbol{w}}\|_1 R\sqrt{\frac{2\log(2d)}{n}} + \sqrt{\frac{2\ln(2/\tau)}{n}}$$

for all $h \in \hat{\mathcal{H}}_\varepsilon$, with probability at least $1 - \tau$ over the choice of training sample, i.e. our choice of constraints allow the generalization error upper bound to increase by $2\delta\|\tilde{\boldsymbol{w}}\|_1 R\sqrt{\frac{2\log(2d)}{n}}$.

3.4. Solution via Linear Programs

In this section, we show how to calculate minimum and maximum relevance bounds using linear programs. This not only shows that our method is easy to implement, but also that the bounds defined by $\text{minRel}((x_i, y_i)_{i=1}^n, j)$ and $\text{maxRel}((x_i, y_i)_{i=1}^n, j)$ are unique and can be calculated in polynomial time. Proofs of the asserted equivalences can be found in Appendix A.

Theorem 2. $\mathrm{minRel}((x_i, y_i)_{i=1}^n, j)$ *is equivalent to the linear program*

$$\begin{aligned}\mathrm{minLP}((x_i, y_i)_{i=1}^n, j) : \min_{\hat{\boldsymbol{w}}, \boldsymbol{w}, b, \boldsymbol{\xi}} \ & \hat{w}_j && (2)\\ \text{s.t. } & w_i - \hat{w}_i \le 0,\ -w_i - \hat{w}_i \le 0, \qquad i = 1, \ldots, d && (3)\\ & y_i(\boldsymbol{w}^\top x_i - b) \ge 1 - \xi_i, \xi_i \ge 0, \quad i = 1, \ldots, n && (4)\\ & \sum_{i=1}^d \hat{w}_i \le (1 + \delta) \cdot \mu && (5)\\ & \sum_{i=1}^n \xi_i \le \rho && (6)\end{aligned}$$

and if $(\hat{\boldsymbol{w}}, \boldsymbol{w}, b, \boldsymbol{\xi})$ *is an optimal point of* $\mathrm{minLP}((x_i, y_i)_{i=1}^n, j)$, *then* $(\boldsymbol{w}, b, \boldsymbol{\xi})$ *is an optimal point of* $\mathrm{minRel}((x_i, y_i)_{i=1}^n, j)$.

Essentially, we reformulate the problem $\mathrm{minRel}((x_i, y_i)_{i=1}^n, j)$ by introducing the auxiliary vector $\hat{\boldsymbol{w}}$. The constraints in (3) enforce $|w_i| \le \hat{w}_i$ for all $i = 1, \ldots, d$, so that $\hat{\boldsymbol{w}}$ upper bounds the element-wise absolute value of $\boldsymbol{w}$. This fact is used in constraint (5) to upper bound the l_1-norm of $\boldsymbol{w}$. At the same time, the objective function in (2) encourages $\hat{w}_j$ to be as small as possible, so that $\hat{w}_j = |w_j|$.

Theorem 3. *The maximum relevance bound is equivalent to taking the maximum of the optimal values of the linear programs*

$$\begin{aligned}\mathrm{maxLPNeg}((x_i, y_i)_{i=1}^n, j) : & && (7)\\ \max_{\hat{\boldsymbol{w}}, \boldsymbol{w}, b, \boldsymbol{\xi}} \ & \hat{w}_j && (8)\\ \text{s.t. } & w_i - \hat{w}_i \le 0, -w_i - \hat{w}_i \le 0, \qquad i = 1, \ldots, d && (9)\\ & \hat{w}_j + w_j \le 0 && (10)\\ & y_i(\boldsymbol{w}^\top x_i - b) \ge 1 - \xi_i, \xi_i \ge 0, \quad i = 1, \ldots, n && (11)\\ & \sum_{i=1}^d \hat{w}_i \le (1 + \delta) \cdot \mu && (12)\\ & \sum_{i=1}^n \xi_i \le \rho && (13)\end{aligned}$$

and

$$\text{maxLPPos}((x_i, y_i)_{i=1}^n, j) : \tag{14}$$

$$\max_{\hat{\boldsymbol{w}}, \boldsymbol{w}, b, \boldsymbol{\xi}} \hat{w}_j \tag{15}$$

$$\text{s. t.} w_i - \hat{w}_i \leq 0, -w_i - \hat{w}_i \leq 0, \qquad i = 1, \dots, d \tag{16}$$

$$\hat{w}_j - w_j \leq 0 \tag{17}$$

$$y_i(\boldsymbol{w}^\top x_i - b) \geq 1 - \boldsymbol{\xi} i, \boldsymbol{\xi} i \geq 0, \quad i = 1, \dots, n \tag{18}$$

$$\sum_{i=1}^d \hat{w}_i \leq (1 + \delta) \cdot \mu \tag{19}$$

$$\sum_{i=1}^n \xi_i \leq \rho \tag{20}$$

That is: If $(\hat{\boldsymbol{w}}^+, \boldsymbol{w}^+, b^+, \boldsymbol{\xi}^+)$ *is an optimal point of* $\text{maxLPPos}((x_i, y_i)_{i=1}^n, j)$ *and* $(\hat{\boldsymbol{w}}^-, \boldsymbol{w}^-, b^-, \boldsymbol{\xi}^-)$ *is an optimal point of* $\text{maxLPNeg}((x_i, y_i)_{i=1}^n, j)$*, then*

$$(\boldsymbol{w}^x, b^x, \boldsymbol{\xi}^x) : x \in \arg\min_{\{+,-\}} \{\hat{w}_j^+, \hat{w}_j^-\}$$

is an optimal point of $\text{maxRel}((x_i, y_i)_{i=1}^n, j)$*.*

Reformulating $\text{maxRel}((x_i, y_i)_{i=1}^n, j)$ as a single linear program is not possible as its objective is to maximize a convex function – the absolute value function. We compensate by dividing the feasible set of $\text{maxRel}((x_i, y_i)_{i=1}^n, j)$ into two parts – one where $w_j \leq 0$ and one where $w_j \geq 0$. This division is enforced by constraints (10) and (17), since $\hat{w}_j \geq 0$ as a consequence of constraints (9) and (16). On the new feasible sets, $|w_j|$ can be written as $-w_j$ and w_j, respectively and optimization via linear programs becomes possible using an auxiliary vector $\hat{\boldsymbol{w}}$ as in Theorem 2.

Using this formulation of the optimization problems as linear programs, our method is easy to implement using any pre-existing SVM and LP solvers. In the following Section, we test its performance on real-world and toy datasets.

4. Experiments

In the following, we show how our method, which we dub FeReL (**Fe**ature **Re**levance for **L**inear Classification), performs on a several of datasets, both synthetic and from the biomedical domain. We have made the Python implementation of Ferel used for these experiments available online.[2]

[2] https://github.com/cgoepfert/ferel

Table 1: Our two data settings. They differ in sample size (size) and the number of strongly relevant (str.), weakly relevant (weak.), and irrelevant features (irrel.).

	str.	weak.	irrel.	size
Setting A	1	2	11	512
Setting B	6	6	6	256

4.1. Comparison to other methods on data with known ground truth

In order to test our method in situations with known ground truth, we create two synthetic data sets with new configurations of strongly relevant, weakly relevant, and irrelevant features, as well as new sample sizes as compared to our original analysis [27]. The objective in each case is the identification of the all-relevant feature set, that is, of all strongly and weakly relevant features. We compare the results of our method to those of feature selection via an l_2-regularized linear classifier (Ridge), an l_1-regularized linear classifier (Lasso), an l_1 and l_2-regularized linear classifier (Elastic Net), and Boruta [17, 18]. For the linear classifiers, a feature is considered relevant if the activation of the corresponding weight in the normal vector to the separating hyperplane is above 10^{-5}. For Boruta, we used the Python implementation `boruta_py` available online[3].

Our method (Ferel) considers a feature as relevant if its maximum relevance bound is above 10^{-5}. Hyperparameters were tuned using 10-fold cross-validation.

The data sets are created according to two different randomized settings which are summarized in Table 1. Here, we create two weakly relevant features by duplicating a single strongly relevant feature, thus creating features that are informative but redundant.

For each setting, we average precision, recall and F1-measure over 10 random instances. Reported precision and recall refer to the comparison of the selected feature sets to the (known) set of all relevant features. Setting A simulates a situation where most of the observed features are irrelevant to the hypothesis class, which can cause the performance of some classifiers to degrade, but is not uncommon in an explorative setting. Setting B simulates a balanced situation where the solutions of the minimal-optimal problem differ markedly from the all-relevant solution. The results can be found in Table 2. Ferel achieves the highest F1-score in both settings. Interestingly, across all methods tested, the recall is quite high while precision tends to be low. This means that even Lasso, which should in theory select only a subset of weakly relevant features, selects all strongly and weakly relevant features – but then selects several irrelevant

[3] `www.github.com/scikit-learn-contrib/boruta_py`

Table 2: Averaged results for Setting A and Setting B.

	Setting A			Setting B		
	Precision	Recall	F1	Precision	Recall	F1
Ridge	0.21	1.00	0.35	0.67	1.00	0.80
Lasso	0.96	1.00	0.98	0.71	1.00	0.83
Elastic Net	0.40	1.00	0.56	0.90	1.00	0.94
Boruta	0.82	1.00	0.89	0.99	0.93	0.96
Ferel	1.00	1.00	**1.00**	0.98	0.98	**0.98**

features as well. The worst precision across both settings is demonstrated by l_2-regularized SVM. This is understandable since the l_2-regularization does not enforce any type of feature selection. Lasso shows a much higher precision in Setting A, showing that it is not as adversely affected by the high number of irrelevant features present. However, its precision drops almost to the level of Ridge in Setting B, where the number of samples is lower, while the number of features is higher. Ferel, which shows perfect performance in Setting A, also shows a small decline in performance in Setting B, where it is overtaken by Boruta concerning precision. The different qualities of results between Settings A and B show that a detailed analysis of the behavior of our and other feature selection methods under low sample sizes is of interest. A study of the behavior of Ferel on one such data set can be found in the following Section.

4.2. Adrenal Dataset

In our previous work [27], we used Ferel to perform an analysis of the adrenal gland metabolomics dataset, which has been described by Biehl et al. [28]. It consists of 147 data points corresponding to adrenocortical carcinoma or adenoma, respectively, described by 32 steroid markers which relate to five different regimes of the underlying metabolic processes. As is common in this type of application, the data dimensionality is relatively high compared to the size of the data set. We are therefore interested in the stability of our method across different train-test splits. We have analyzed the stability by calculating the standard deviation of the achieved minimum and maximum relevance bounds as compared to the standard deviation of the entries of the baseline classifier across 64 90-10 train-test-splits. The results are given in Figure 2. We observe that the ratio of standard deviations is close to 1 for most features, which shows that our method does not introduce significant instability in these cases. A comparatively large increase of standard deviation can be observed for features 2, 9 and 31, which are considered irrelevant both by the baseline classifier and by Ferel.

Figure 3 shows the mean minimum and maximum relevance bounds averaged over all train-test-splits.

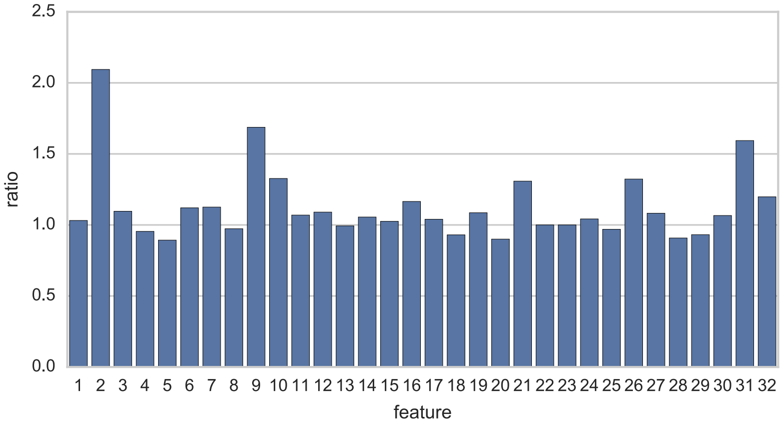

Figure 2: Per-feature ratio of the standard deviation of the maximum relevance bounds found by our method to the corresponding weight in the baseline classifier.

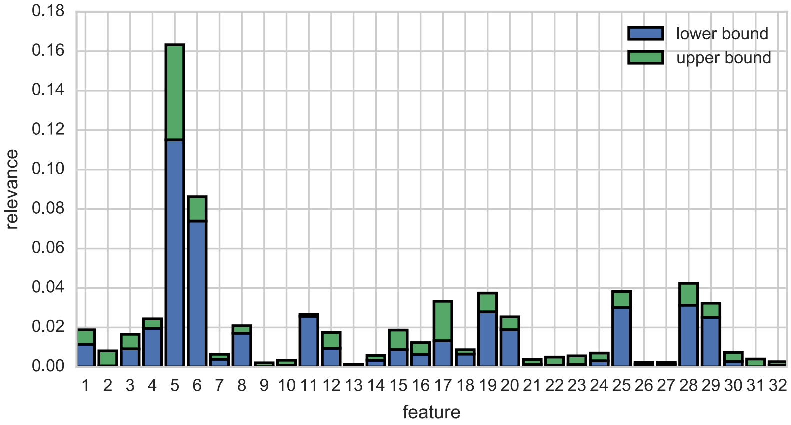

Figure 3: Results of Ferel on the adrenal dataset, averaged over 64 train-test splits.

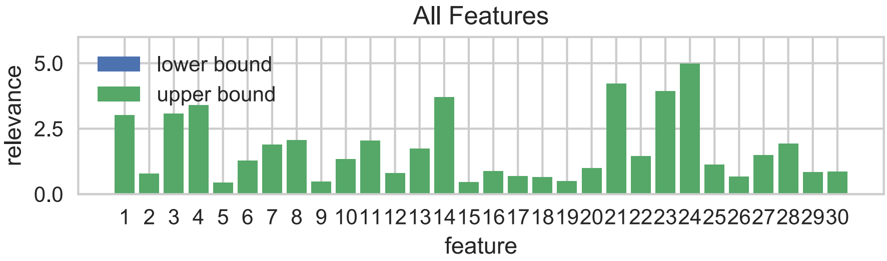

Figure 4: Results of Ferel on the Wisconsin Breast Cancer (Diagnostic) dataset using all features. The lower bound is 0 for all features.

4.3. Breast Cancer Wisconsin (Diagnostic) Dataset

We use Ferel to examine the Wisconsin Breast Cancer (Diagnostic) [29] dataset from [30], which contains properties of cell nuclei from malignant and benign tissue samples in the form of 30 features. The resulting feature relevance profile is given in Figure 4. It indicates that every single feature is weakly relevant and so, no feature is irrelevant – maybe more importantly, no feature is strongly relevant, which suggests that any one of them could be discarded without adversely affecting classification. We validated this empirically by training a classifier on all feature subsets with 29 features. Performance was not worse than with all 30 features. The observed relevances are not surprising, considering the relations between many of the dataset's features. They are ten triplets comprising *mean*, *standard error* and *worst* of certain features and include the radius, perimeter and area of certain structures. When we run Ferel on reduced versions of the dataset that contain only mean, standard error or worst features (see Figure 5), we see that all of the *mean* features remain weakly relevant. Considering only the *standard error* features, the 14th feature becomes strongly relevant. It describes the *standard error in area* covered by the nuclei. Of the *worst* features, the 22nd feature is considered strongly relevant. This feature is the *worst texture*, a number that describes how irregular a nucleus's color is. Compared to training with all features, the F1-score drops from 0.9589 to 0.9577, 0.8889, and 0.9429, respectively. These experiments demonstrate the benefits of our relevance taxonomy: We clearly observe features that are redundant in the presence of other features become indispensable when some of the other features are removed. The similar performance of the *mean* and *worst* feature subsets suggest that both could contain minimal optimal sets. However, the cost of measuring and recording the features may vary greatly between both sets, so that simply identifying one minimal optimal set is far from ideal.

5. Conclusion

We have defined and tackled the specific all-relevant feature selection problem for the hypothesis class of linear classifiers, stating it as the problem of

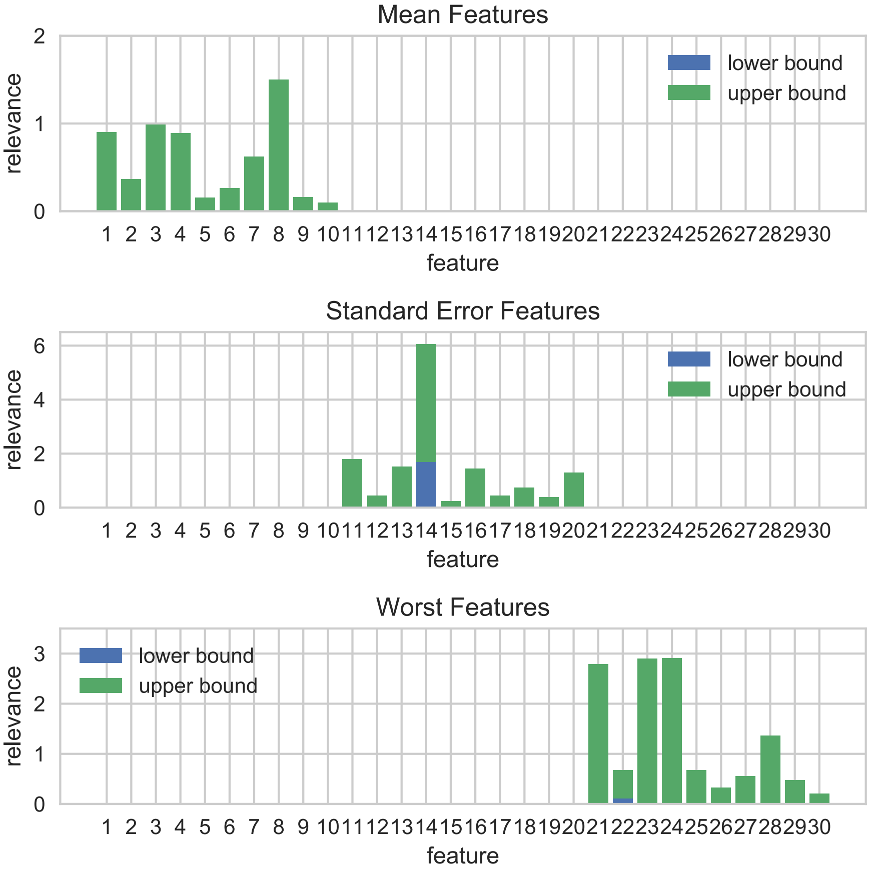

Figure 5: Results of Ferel on the Wisconsin Breast Cancer (Diagnostic) dataset. In the first, second, and third plot, only the mean, standard error and worst features are used, respectively – the remaining features are plotted for easier comparison with their lower and upper bounds set to 0.

finding minimum and maximum relevance bounds in the class of all equivalent hypotheses as concerns hinge loss and l_1-norm of the weight vector. We have argued that this approach constitutes a sensible approximation of the specific all-relevant problem, an approximation that is necessary as we do not have access to the underlying data distribution. As an added benefit, our method allows for the distinction between strongly and weakly relevant features, which is not required as part of the specific all-relevant problem, but nonetheless provides valuable information for practitioners. Furthermore, we have shown that the necessary search over the set of linear classifiers considered equivalent can be efficiently performed using linear programs, which yield unique results in polynomial time.

We have augmented our previous analyses by comparing our method with two other all-relevant feature selection methods on new configurations of synthetic data with known ground truth, with our method outperforming both. In addition, we have tested our method on an additional real-world data set and analyzed the stability of our method on real-world data over repeated train-test splits. This is an important concern due to the typically high dimensionality of data from the biomedical application domain as compared to data set size.

In practice, the proposed method opens a way for an intelligent and interactive analysis of linear models based on all possibly relevant features for a classification problem, thus facilitating data introspection as well as classifier design. Additionally, the framework we have developed for tackling the specific all-relevant problem for a linear hypothesis class is transferable to other hypothesis classes or other performance measures, such as area-under-the-curve instead of generalization error. Area under the curve evaluation and optimization is particularly useful for imbalanced classes, as is common in the biomedical domain where the number of healthy patients typically heavily outweighs the number of sick patients. In the future, we will tackle these types of extensions, as well as developing methods that automatically visualize the mutual relationships of weakly relevant features.

Acknowledgement

Funding by the CITEC cluster of excellence and by the German-Canadian DFG International Research Training Group Didy (1906/1) is gratefully acknowledged.

References

[1] R. Kohavi, G. H. John, Wrappers for Feature Subset Selection, Artif. Intell. 97 (1-2) (1997) 273–324.

[2] I. Guyon, A. Elisseeff, An Introduction to Variable and Feature Selection, Journal of Machine Learning Research 3 (2003) 1157–1182.

[3] M. T. Ribeiro, S. Singh, C. Guestrin, "why should i trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, ACM, 2016, pp. 1135–1144.

[4] V. Van Belle, P. Lisboa, White box radial basis function classifiers with component selection for clinical prediction models, Artificial Intelligence in Medicine 60 (1) (2014) 53–64.

[5] G. Bhanot, M. Biehl, T. Villmann, D. Zühlke, Integration of Expert Knowledge for Interpretable Models in Biomedical Data Analysis (Dagstuhl Seminar 16261).

[6] A. Nguyen, J. Yosinski, J. Clune, Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images, ArXiv e-printsarXiv:1412.1897.

[7] J. Tang, S. Alelyani, H. Liu, Feature Selection for Classification: A Review, Chapman and Hall/CRC, 2014, pp. 37–64.

[8] J. H. Gennari, P. Langley, D. Fisher, Models of incremental concept formation, Artificial Intelligence 40 (1) (1989) 11–61.

[9] I. M. Guyon, S. R. Gunn, A. Ben-Hur, G. Dror, Result Analysis of the NIPS 2003 Feature Selection Challenge, advances in Neural Information Processing Systems (2004).

[10] F. Coelho, A. de Pádua Braga, M. Verleysen, A mutual information estimator for continuous and discrete variables applied to feature selection and classification problems, Int. J. Computational Intelligence Systems 9 (4) (2016) 726–733.

[11] R. Tibshirani, Regression Shrinkage and Selection Via the Lasso, Journal of the Royal Statistical Society, Series B 58 (1994) 267–288.

[12] J. Huang, P. Breheny, S. Ma, A selective review of group selection in high-dimensional models, Statistical science : a review journal of the Institute of Mathematical Statistics 27 (4).

[13] D. M. Witten, A. Shojaie, F. Zhang, The cluster elastic net for high-dimensional regression with unknown variable grouping, Technometrics : a journal of statistics for the physical, chemical, and engineering sciences 56 (1) 112–122.

[14] H. Zou, T. Hastie, Regularization and Variable Selection via the Elastic Net, Journal of the Royal Statistical Society. Series B (Statistical Methodology) 67 (2) (2005) 301–320.

[15] L. Yu, H. Liu, Efficient Feature Selection via Analysis of Relevance and Redundancy, Journal of Machine Learning Research 5 (2004) 1205–1224.

[16] P. Schneider, M. Biehl, B. Hammer, Adaptive relevance matrices in learning vector quantization, Neural Computation 21 (12) (2009) 3532–3561.

[17] W. R. Rudnicki, M. Wrzesień, W. Paja, All Relevant Feature Selection Methods and Applications, in: U. Stańczyk, L. C. Jain (Eds.), Feature Selection for Data and Pattern Recognition, no. 584 in Studies in Computational Intelligence, Springer Berlin Heidelberg, 2015, pp. 11–28.

[18] M. B. Kursa, W. R. Rudnicki, The All Relevant Feature Selection using Random Forest, ArXiv e-printsarXiv:1106.5112.

[19] Frénay, B., D. Hofmann, A. Schulz, M. Biehl, B. Hammer, Valid interpretation of feature relevance for linear data mappings, in: 2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), 2014, pp. 149–156.

[20] A. Schulz, B. Mokbel, M. Biehl, B. Hammer, Inferring Feature Relevances From Metric Learning, in: 2015 IEEE Symposium Series on Computational Intelligence, 2015, pp. 1599–1606.

[21] J. Jia, B. Yu, On model selection consistency of the elastic net when p » n, Statistica Sinica 20 (2) (2010) 595–611.

[22] R. Nilsson, J. M. Peña, J. Björkegren, J. Tegnér, Consistent Feature Selection for Pattern Recognition in Polynomial Time, Journal of Machine Learning Research 8 (2007) 589–612.

[23] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene Selection for Cancer Classification using Support Vector Machines, Machine Learning 46 (1-3) (2002) 389–422.

[24] Z. Tang, Y. Shen, X. Zhang, N. Yi, The Spike-and-Slab Lasso Generalized Linear Models for Prediction and Associated Genes Detection, Genetics (2017) 77–88.

[25] E. Tuv, A. Borisov, K. Torkkola, Feature Selection Using Ensemble Based Ranking Against Artificial Contrasts, in: The 2006 IEEE International Joint Conference on Neural Network Proceedings, 2006, pp. 2181–2186.

[26] S. Shalev-Shwartz, S. Ben-David, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014.

[27] C. Göpfert, L. Pfannschmidt, B. Hammer, Feature Relevance Bounds for Linear Classification, in: 25th European Symposium on Artificial Neural Networks (ESANN), 2017.

[28] M. Biehl, P. Schneider, D. J. Smith, H. Stiekema, A. E. Taylor, B. A. Hughes, C. H. L. Shackleton, P. M. Stewart, W. Arlt, Matrix relevance LVQ in steroid metabolomics based classification of adrenal tumors, in: 20th European Symposium on Artificial Neural Networks (ESANN), 2012, pp. 423–428.

[29] W. N. Street, W. H. Wolberg, O. L. Mangasarian, Nuclear feature extraction for breast tumor diagnosis, in: Biomedical Image Processing and Biomedical Visualization, Vol. 1905, International Society for Optics and Photonics, 1993, pp. 861–871.

[30] M. Lichman, UCI machine learning repository (2013).
URL http://archive.ics.uci.edu/ml

Appendix A. Proofs of Theorem 2 and 3

First, we prove a support theorem that we will use in the proofs of Theorems 2 and 3:

Theorem 4. *Regard the two optimization problems*

$$\textit{Problem 1:}\ \min_x \mathrm{h}_1(x)\,\mathrm{s.\,t.}\, x \in A_1$$

and

$$\textit{Problem 2:}\ \min_x \mathrm{h}_2(x)\,\mathrm{s.\,t.}\, x \in A_2 \tag{A.1}$$

If there exist maps $f : A_1 \to A_2$ and $g : A_2 \to A_1$ such that for all $x \in A_1$, $y \in A_2$:

$$\mathrm{h}_2(y) < \mathrm{h}_2(f(x)) \Rightarrow \mathrm{h}_1(g(y)) < \mathrm{h}_1(x) \tag{A.2}$$
$$\mathrm{h}_1(x) < \mathrm{h}_1(g(y)) \Rightarrow \mathrm{h}_2(f(x)) < \mathrm{h}_2(y), \tag{A.3}$$

then Problems 1 and 2 are equivalent, that is, one can easily be solved by solving the other.

Proof. Let x_{opt} be an optimal point of Problem 1. Then $\mathrm{h}_2(f(x_{opt})) \leq \mathrm{h}_2(y)$ for all $y \in A_2$, i.e. $f(x_{opt})$ is an optimal point of Problem 2, because $\mathrm{h}_2(y) < \mathrm{h}_2(f(x_{opt}))$ would imply $\mathrm{h}_1(x_{opt}) > \mathrm{h}_1(g(y))$ according to (A.2), which contradicts the optimality of x_{opt}. Switching the roles of Problem 1 and Problem 2 shows that if y_{opt} is an optimal point of Problem 2, $g(y_{opt})$ is an optimal point for Problem 1. □

Now, we will define the mappings f and g: For brevity, we will suppress the arguments $((x_i, y_i)_{i=1}^n, j)$ to the optimization problems and refer to them by name only. The domain of minRel and maxRel is $\mathbb{R}^{d+1+n}$. Their feasible sets are identical and denoted by A. The domain of minLP, maxLPPos, and maxLPNeg is $\mathbb{R}^{d+d+1+n}$ and we denote their feasible sets by B_{min}, B^+_{max} and B^-_{max}, respectively. The mappings f and g are defined by

$$f : \mathbb{R}^{d+1+n} \to \mathbb{R}^{d+d+1+n}$$
$$(\boldsymbol{w}, b, \boldsymbol{\xi}) \mapsto (|\boldsymbol{w}|, \boldsymbol{w}, b, \boldsymbol{\xi})$$

and

$$g : \mathbb{R}^{d+d+1+n} \to \mathbb{R}^{d+1+n}$$
$$(\hat{\boldsymbol{w}}, \boldsymbol{w}, b, \boldsymbol{\xi}) \mapsto (\boldsymbol{w}, b, \boldsymbol{\xi})$$

Let $A^+ = \{(\boldsymbol{w}, b, \boldsymbol{\xi}) \in A \mid w_j \geq 0\}$ and $A^- = \{(\boldsymbol{w}, b, \boldsymbol{\xi}) \in A \mid w_j \leq 0\}$. Clearly, if $(\boldsymbol{w}, b, \boldsymbol{\xi}) \in A$, A^+ or A^-, then $f(\boldsymbol{w}, b, \boldsymbol{\xi}) \in B_{min}$, B^+_{max} or B^-_{max}, respectively and vice versa. Thus, f and g are transformations between the feasible sets of minRel and minLP. In the Proof of Theorem 3, we will introduce optimization problems with feasible sets A^+ and A^- that can be combined to solve maxRel. Then, it only remains to show that (A.2) and (A.3) hold in each case.

Proof of Theorem 2. The objective function of minRel is

$$\mathrm{h}_1(\boldsymbol{w}, b, \boldsymbol{\xi}) = |w_j|$$

and the objective function of minLP is

$$\mathrm{h}_2(\hat{\boldsymbol{w}}, \boldsymbol{w}, b, \boldsymbol{\xi}) = \hat{w}_j.$$

Let $(\boldsymbol{w}, b, \boldsymbol{\xi}) \in A$ and $(\hat{\boldsymbol{w}}, \boldsymbol{w}', b', \boldsymbol{\xi}') \in B_{min}$. Then, per definition,

$$\mathrm{h}_2(\hat{\boldsymbol{w}}, \boldsymbol{w}', b', \boldsymbol{\xi}') < \mathrm{h}_2(f(\boldsymbol{w}, b, \boldsymbol{\xi})) \Leftrightarrow \hat{w}_j < |w_j|$$

which implies $|w'_j| < |w_j|$ due to (3), so that $\mathrm{h}_1(g(\hat{\boldsymbol{w}}, \boldsymbol{w}', b', \boldsymbol{\xi}')) < \mathrm{h}_1(\boldsymbol{w}, b, \boldsymbol{\xi})$.

On the other hand,

$$\mathrm{h}_1(\boldsymbol{w}, b, \boldsymbol{\xi}) < \mathrm{h}_1(g(\hat{\boldsymbol{w}}, \boldsymbol{w}', b', \boldsymbol{\xi}')) \Leftrightarrow |w_j| < |w'_j|$$

which by (3) implies $|w_j| < \hat{w}_j$, so that $\mathrm{h}_2(f(\boldsymbol{w}, b, \boldsymbol{\xi})) < \mathrm{h}_2(\hat{\boldsymbol{w}}, \boldsymbol{w}', b', \boldsymbol{\xi}')$. □

Proof of Theorem 3. Regard the two problems

$$\mathrm{maxRelPos}((x_i, y_i)_{i=1}^n, j) : \min_{\boldsymbol{w}, b, \boldsymbol{\xi}} \; -|w_j| \,\mathrm{s.\,t.} (\boldsymbol{w}, b, \boldsymbol{\xi}) \in A^+$$

and

$$\mathrm{maxRelNeg}((x_i, y_i)_{i=1}^n, j) : \min_{\boldsymbol{w}, b, \boldsymbol{\xi}} \; -|w_j| \,\mathrm{s.\,t.} (\boldsymbol{w}, b, \boldsymbol{\xi}) \in A^-$$

Since the objective functions of maxRel, maxRelPos and maxRelNeg are identical, and the union of the feasible sets of maxRelPos and maxRelNeg is the feasible set of maxRel, maxRel can be solved by solving maxRelPos and maxRelNeg, and taking the result that gives the higher value of $|w_j|$. It remains to show that maxRelPos is equivalent to maxLPPos and maxRelNeg is equivalent to maxLPNeg. We will prove the first equivalence.

The objective function of maxRelPos is

$$\mathrm{h}_1(\boldsymbol{w}, b, \boldsymbol{\xi}) = -|w_j|$$

and the objective function of maxLPPos is

$$\mathrm{h}_2(\hat{\boldsymbol{w}}, \boldsymbol{w}, b, \boldsymbol{\xi}) = -\hat{w}_j.$$

Let $(\boldsymbol{w}, b, \boldsymbol{\xi}) \in A$ and $(\hat{\boldsymbol{w}}, \boldsymbol{w}', b', \boldsymbol{\xi}') \in B_{min}$. Then,

$$\mathrm{h}_2(\hat{\boldsymbol{w}}, \boldsymbol{w}', b', \boldsymbol{\xi}') < \mathrm{h}_2(f(\boldsymbol{w}, b, \boldsymbol{\xi})) \Leftrightarrow -\hat{w}_j < -|w_j|.$$

Since $-w_j' \leq -\hat{w}_j$ by (17), this implies $-w_j' < -|w_j|$, and because $w_j' \geq 0$ by (16) and (17), we have $-|w_j'| < -|w_j|$. This shows that $\mathrm{h}_1(g(\hat{\boldsymbol{w}}, \boldsymbol{w}', b', \boldsymbol{\xi}')) < \mathrm{h}_1(\boldsymbol{w}, b, \boldsymbol{\xi})$.

On the other hand,

$$\mathrm{h}_1(\boldsymbol{w}, b, \boldsymbol{\xi}) < \mathrm{h}_1(g(\hat{\boldsymbol{w}}, \boldsymbol{w}', b', \boldsymbol{\xi}')) \Leftrightarrow -|w_j| < -|w_j'|.$$

Since $w_j' \geq 0$ by (16) and (17), this implies $-|w_j| < -w_j'$, and because $-w_j' \leq -\hat{w}_j$ by (17), we have $-|w_j| < -\hat{w}_j$. This shows that $\mathrm{h}_2(f(\boldsymbol{w}, b, \boldsymbol{\xi})) < \mathrm{h}_2(\hat{\boldsymbol{w}}, \boldsymbol{w}, b, \boldsymbol{\xi})$.

The proof of equivalence of maxRelNeg and maxLPNeg uses the same arguments, with $w_j' \leq -\hat{w}_j$ instead of $-w_j' \leq -\hat{w}_j$ and $w_j' \leq 0$ instead of $w_j' \geq 0$. □