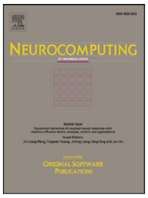
Neurocomputing 000 (2018) 1–11

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



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

ARTICLE INFO

Article history:
Received 9 July 2017
Revised 5 November 2017
Accepted 26 November 2017
Available online xxx

Keywords:
Feature relevance
Feature selection
Interpretability
All-relevant
Linear classification

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.

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 [3–5], in particular in the light of easily fooled machine learning models [6]. Despite 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 et al. [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 [1,10–16].

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 that 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 that 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.

* Corresponding author.
E-mail addresses: cgoepfert@techfak.uni-bielefeld.de (C. Göpfert), lpfannschmidt@techfak.uni-bielefeld.de (L. Pfannschmidt), jgoepfert@techfak.uni-bielefeld.de (J.P. Göpfert), bhammer@techfak.uni-bielefeld.de (B. Hammer).

https://doi.org/10.1016/j.neucom.2017.11.074
0925-2312/© 2018 Elsevier B.V. All rights reserved.

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 that 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 that 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 I* that maps a training sample to some *classification rule* or *hypothesis* $h : \mathbb{R}^d \to \{-1, 1\}$ whereby the set Im(I) of classification rules the inducer can map to is called the *hypothesis space* $\mathcal{H}$ of *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

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