Feature Relevance Bounds for Linear Classification

Christina Göpfert, Lukas Pfannschmidt and Barbara Hammer *

CITEC center of excellence
Bielefeld University - Germany

(This is a preprint of the publication [1], as provided by the authors.)

Abstract

Biomedical applications often aim for an identification of relevant features for a given classification task, since these carry the promise of semantic insight into the underlying process. For correlated input dimensions, feature relevances are not unique, and the identification of meaningful subtle biomarkers remains a challenge. One approach is to identify *intervals for the possible relevance* of given features, a problem related to *all-relevant* feature determination. In this contribution, we address the important case of linear classifiers and we reformulate the inference of feature relevance bounds as a convex optimization problem. We demonstrate the superiority of the resulting technique in comparison to popular feature-relevance determination methods in several benchmarks.

1 Introduction

The increase in data availability in the biomedical domain has led to growing opportunities for machine learning applications. Besides mere statistical inference, model interpretability offers one possibility to gain insight into the underlying processes and to align models and expert knowledge [2, 3]. One popular form of model interpretability is given by feature relevance determination or selection schemes, which enable users to identify the most relevant input variables as potential biomarkers. Successful applications can be based on metric learning or sparse linear models, as in [4, 5, 6, 7].

Feature selection focuses on algorithms that identify relevant features for machine learning tasks. Integrated techniques such as sparse linear models or relevance learning combine the benefit of computational efficiency with a natural treatment of multivariate feature relevance [8, 9, 10]. In particular for high dimensional data, the result is not unique, which can be attributed to the presence of redundant (weakly relevant) features [11]. As recently demonstrated

*Funding within the DFG international research training group DiDy (IGK 1906) and the CITEC center of excellence (EXC 277) is gratefully acknowledged.

in [12, 13], raw feature relevance profiles can be misleading in such settings, and discretion is needed to extract meaningful feature subsets. There exists a variety of methods to identify minimal feature subsets, whereby ambiguities are mostly resolved randomly and subtle signals are usually neglected. Contrarily, the all-relevant problem aims for *all* potentially relevant features. This enables a practitioner to choose the best biomarkers for a given setting interactively.

The all-relevant feature selection problem is provably more difficult than identification of only strongly relevant features or a minimal feature subset, and only few methods tackle it so far [14]. One possible all-relevant feature selection method is the Elastic Net, which enforces sparsity and encourages grouping by combining L_1- and L_2-penalties [15]. Another option is Boruta [16], which calculates an importance measure based on random forests and determines relevance by its comparison to artificial contrast variables. However, to the best of our knowledge, no approach addresses a weighting of strongly and weakly relevant features for a given linear classification by means of linear programs.

In the following, we state the problem of determining feature relevance bounds for a linear classification task in terms of linear programs yielding unique feature relevance intervals, and we discuss how to extract strongly and weakly relevant features for linear dependencies based thereon. We show that the results are superior to alternative schemes including Boruta, L_1-constrained SVM, and Elastic Net on benchmark data with known ground truth, and we demonstrate the applicability for two examples from the biomedical domain.

2 Relevance bounds for feature selection

Given a binary classification problem represented by labeled data points $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \{-1, 1\}$, $i = 1, \ldots, n$, our goal is to assess the relevance of each feature for linear classification. Kohavi and John [11] distinguish between three different levels of relevance: A feature is *strongly relevant* if its removal lowers the performance of the optimal Bayes classifier; it is *weakly relevant* if it is not strongly relevant but there exists a subset of features such that it is strongly relevant among those, and it is *irrelevant* if it is neither strongly nor weakly relevant. Inspired by this taxonomy, we investigate feature relevance for the important case of linear classification. Clearly, more than a single importance value for each feature is needed to distinguish between both strong and weak relevance, and weak relevance and irrelevance. Thus, we aim to determine the minimal and maximal relevance of each feature *taking into account the potential influence of all other features*. If the minimal relevance of a feature is greater than zero, it is strongly relevant. If its maximal relevance is zero, it is irrelevant. If the lower bound is zero, and the upper bound greater than zero, it is weakly relevant.

For linear classifiers, the absolute values of the weight vector that defines a separating hyperplane can be taken as an indicator of feature relevance [17]. When this weight vector is computed using L_2-regularization, highly correlated features share their weight, and groups of weakly relevant features may be mistaken for noise. L_1-regularization enforces a sparse weight vector, revealing the potential importance of single weakly or strongly relevant features, but not of all of those. We also use L_1-regularization, as it permits weight to be shifted

within a group of weakly relevant features, but, mimicking the idea proposed in [12, 13], we use a set of optimization problems to reveal the relevance bounds.

In the following, let $(\tilde{\boldsymbol{w}}, \tilde{b}, \tilde{\boldsymbol{\xi}})$ denote the solution of a linear SVM with regularization C, where $\tilde{\boldsymbol{\xi}} = (\tilde{\xi}_1, \ldots, \tilde{\xi}_n)$ are slack variables controlling margin intrusion:
$\min_{\tilde{\boldsymbol{w}}, \tilde{b}, \tilde{\boldsymbol{\xi}}} \|\tilde{\boldsymbol{w}}\|_2 + C \cdot \sum_{i=1}^n \tilde{\xi}_i$ s. t. $y_i(\tilde{\boldsymbol{w}} \cdot \boldsymbol{x}_i^\top - \tilde{b}) \geq 1 - \tilde{\xi}_i, \tilde{\xi}_i \geq 0$, $i = 1, \ldots, n$.
The **minimum linear relevance bound** for feature j is defined as:

$$\textbf{Problem I}: \quad \begin{array}{ll} \min_{\boldsymbol{w}, b, \boldsymbol{\xi}} & |w_j| \\ \text{s. t.} & y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i^\top - b) \geq 1 - \xi_i, \xi_i \geq 0, \quad i = 1, \ldots, n \\ & \|\boldsymbol{w}\|_1 + C \cdot \sum_{i=1}^n \xi_i \leq \|\tilde{\boldsymbol{w}}\|_1 + C \cdot \sum_{i=1}^n \tilde{\xi}_i. \end{array}$$

The **maximum linear relevance bound** (**Problem II**) of j, is defined by replacing $\min_{\boldsymbol{w}, b, \boldsymbol{\xi}}$ with $\max_{\boldsymbol{w}, b, \boldsymbol{\xi}}$. Note that the L_1-bound constraint restricts the margin of each candidate hyperplane to at least $1/\sqrt{d}$ times the margin of the original SVM. This factor is minimal as to allow d identical features to concentrate their formerly distributed relevance onto a single feature.

3 Efficient Realization by Linear Programming

Problems I and II can be solved efficiently using linear programs (LP). Here we omit the proofs of equivalence due to space limitations.

Theorem 1. *Problem I is convex and an optimal solution is obtained via the following linear problem with $2d + n + 1$ variables and $2d + n + 1$ constraints:*

$$\begin{array}{ll} \min_{\hat{\boldsymbol{w}}, \boldsymbol{w}, b, \boldsymbol{\xi}} & \hat{w}_j \\ \text{s. t.} & w_i - \hat{w}_i \leq 0, \; -w_i - \hat{w}_i \leq 0, \quad i = 1, \ldots, d \\ & -y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i^\top - b) \leq \xi_i - 1, \qquad i = 1, \ldots, n \\ & \sum_{i=1}^d \hat{w}_i + C \cdot \sum_{i=1}^n \xi_i \leq \mu, \end{array}$$

where $\mu = \|\tilde{\boldsymbol{w}}\|_1 + C \cdot \sum_{i=1}^n \tilde{\xi}_i$. Its optimal solution $(\hat{\boldsymbol{w}}, \boldsymbol{w}, b, \boldsymbol{\xi})$ induces an optimal solution $(\boldsymbol{w}, b, \boldsymbol{\xi})$ of Problem I; it holds $\hat{\boldsymbol{w}} = |\boldsymbol{w}|$.

While Theorem 1 relies on a classical transformation, an LP formalization of Problem II requires a problem specific transformation:

Theorem 2. *Regard the linear programs*

$$\textbf{(a)}: \quad \begin{array}{ll} \max_{\hat{\boldsymbol{w}}, \boldsymbol{w}, b, \boldsymbol{\xi}} & \hat{w}_j \\ \text{s. t.} & w_i - \hat{w}_i \leq 0, -w_i - \hat{w}_i \leq 0, \quad i = 1, \ldots, d \\ & \hat{w}_j + w_j \leq 0 \qquad\qquad\qquad\qquad\qquad\qquad (*) \\ & -y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i^\top - b) \leq \xi_i - 1, \qquad i = 1, \ldots, n \\ & \sum_{i=1}^d \hat{w}_i + C \cdot \sum_{i=1}^n \xi_i \leq \mu, \end{array}$$

and ***(b)*** *where the condition $(*)$ is substituted by $\hat{w}_j - w_j \leq 0$. Let $(\hat{\boldsymbol{w}}^a, \boldsymbol{w}^a, b^a, \boldsymbol{\xi}^a)$ and $(\hat{\boldsymbol{w}}^b, \boldsymbol{w}^b, b^b, \boldsymbol{\xi}^b)$ be optimal solutions of* ***(a)*** *and* ***(b)****. Then, $(\boldsymbol{w}^x, b^x, \boldsymbol{\xi}^x)$ such that $\hat{w}_j^x$ is maximal optimally solves Problem II.*

As a consequence, for linear mappings, feature relevance bounds can be efficiently determined and they are unique. The resulting intervals reveal a detailed measure of the feature relevance when taking all possible models with the same classification accuracy and L_1-norm into account. Based on the resulting bounds, we extract both weakly and strongly relevant features for the considered linear classification task: strongly relevant features are those with strictly positive lower bound (they cannot be deleted from the set without sacrificing model accuracy), while weakly relevant features are those with zero minimum relevance bound but strictly positive upper bound (they contribute to at least one, but not all optimal linear models). For an according feature selection, we determine suitable cutoff values via the relevance bounds related to features obtained after a random permutation along the given data column.

4 Experiments

Artificial data: For comparison we created three datasets with known ground truth, containing $n = 150$ samples and $d = 12$ features each. The number of strongly relevant, weakly relevant, and irrelevant features is characterized by the triplets $(6, 0, 6)$ for Data I, $(0, 6, 6)$ for Data II, and $(3, 4, 3)$ for Data III. The relevant feature dimensions determine a hyperplane that defines class assignments. Weakly relevant features are linear combinations of strongly relevant ones. We compare our method to an L_2-regularized SVM (no explicit feature selection), L_1-regularized SVM (aiming for a minimal optimal set), Elastic Net (all relevant features), Boruta (all relevant features) [16], and a forward/backward selection based on classification performance as proposed in [12] (all relevant features). Hyperparameters are optimized via grid search and 5-fold cross validation. Since C controls the sparsity and estimation error of the resulting weight vector, we aim to analyze its regularization path in the future. Features from linear models are ranked based on their importance weights, where the cutoff is set to 10^{-5} for L_1-regularized models, and the mean feature value for L_2-regularized models and elastic net. The results of all methods are displayed in Table 1. The classification performance is 100% accuracy for all methods and data sets. Reported precision and recall refer to the comparison of the selected feature sets to the (known) set of all relevant features. Not all methods address the all relevant features problem; yet, they also partially fail in settings where they should deliver this solution by design, such as L_1-SVM for Data I. The methods for all relevant feature selection, Elastic Net, Boruta, and forward/backward search, often do not deliver optimal results. Conversely, our method provides an F-score of at least 0.97 in all settings. A python-implementation of our method and the code used to generate our artificial datasets can be found at `https://github.com/lpfann/fri`.

Medical data analysis: We evaluate our method for two data sets from the medical domain: The adrenal gland metabolomics dataset has been described in [18]. 147 data points corresponding to adrenocortical carcinoma or adenoma, respectively, are described by steroid markers which relate to five different regimes of the underlying metabolic processes (see Fig. 1). The binary classification problem is solved with F-score 0.98 and standard deviation $0.5 \cdot 10^{-2}$ for all

Table 1: Precision, recall and F1-values of feature selection methods on synthetic datasets with different properties. Values are averaged over 10 random instances of the data sets.

Data	I			II			III		
	prec.	rec.	F1	prec.	rec.	F1	prec.	rec.	F1
L_2-SVM	1.00	0.82	0.89	1.00	0.83	0.90	1.00	0.70	0.82
L_1-SVM	0.56	1.00	0.72	0.57	1.00	0.72	0.72	1.00	0.83
ElasticNet	1.00	0.85	0.92	1.00	0.83	0.90	1.00	0.76	0.85
Boruta	0.94	0.83	0.87	1.00	0.85	0.91	0.96	0.80	0.87
forw./back.	1.00	0.77	0.86	1.00	0.80	0.87	0.85	0.75	0.79
our method	1.00	0.97	0.98	0.95	1.00	0.97	1.00	0.97	0.98

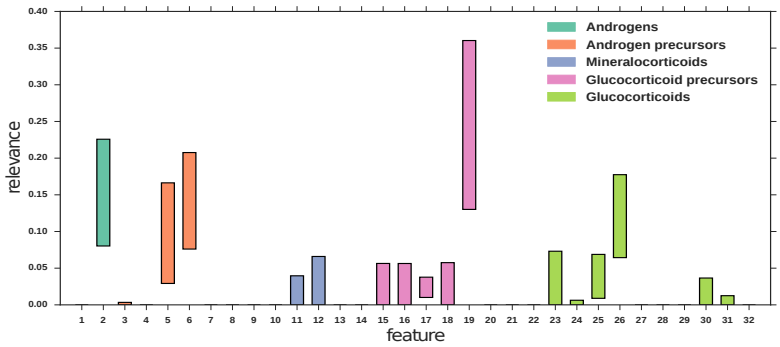

Figure 1: Relevance profile for dataset with features stemming from the grouped cholesterol pathway in the adrenal gland metabolism [18].

models corresponding to minimum/maximum ranks as shown in Fig. 1. Interestingly,we can extract strongly relevant features in each group of the cholesterol pathway except in the *androgen precursors*. The latter is represented by two weakly relevant features whereby their simultaneous removal leads to a degradation of the classification accuracy by 1%. Hence the extracted bounds do not only resemble findings as reported in [18], they also align with prior knowledge about the semantic grouping of underlying metabolic processes. A similar result can be obtained for the Wisconsin diagnostic breast cancer data set [19]. Malignant versus benign samples are predicted based on 30 statistical features which describe the distribution and characteristics of images obtained from a fine needle aspirate. Here the average F-score of the classification result is 0.98 with standard deviation $0.8 \cdot 10^{-3}$. The feature relevance profile as depicted in Fig. 2 singles out a few clear strongly relevant features as well as a handful of weakly relevant ones, which partially directly relate to the underlying semantic correlations of the considered features.

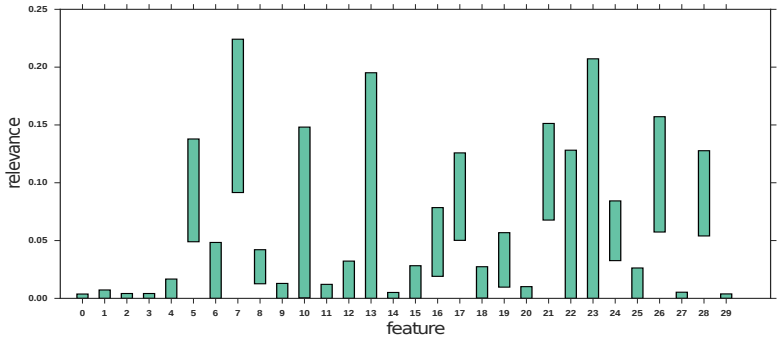

Figure 2: Breast Cancer Wisconsin diagnostic data set ($n = 569, d = 30$, geometric properties of cell imagery); here, features are grouped according to their semantic similarity in blocks of three [19].

5 Conclusion

We have tackled the all-relevant feature selection problem for linear classification, stating it as the problem of finding minimum and maximum relevant bounds in the class of all equivalent models as concerns classification accuracy and L_1-norm. We have transferred this problem to a set of LP problems which yield unique solutions in polynomial time. For artificial data, the technique has proven superior compared to known alternatives, and its results have aligned with prior knowledge on two biomedical problems. In practice, the selection of weakly relevant features for further use depends on the given setting at hand, and the proposed method opens a way for an intelligent interactive analysis based on all possibly relevant biomarker candidates. In the future, we will enhance the model with automatic techniques to also visualize the mutual relationships of weakly relevant features in order to facilitate expert exploration of the results.

References

[1] Christina Göpfert, Lukas Pfannschmidt, and Barbara Hammer. Feature Relevance Bounds for Linear Classification. In *Proceedings of the ESANN*. 25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2017.

[2] Vanya Van Belle and Paulo Lisboa. White box radial basis function classifiers with component selection for clinical prediction models. *Artificial Intelligence in Medicine*, 60(1):53–64, 2014.

[3] Gyan Bhanot, Michael Biehl, Thomas Villmann, and Dietlind Zühlke. Integration of expert knowledge for interpretable models in biomedical data analysis (dagstuhl seminar 16261). *Dagstuhl Reports*, 6(6):88–110, 2016.

[4] Hongbao Cao, Junbo Duan, Dongdong Lin, Yin Yao Shugart, Vince D. Calhoun, and Yu-Ping Wang. Sparse representation based biomarker selection for schizophrenia with integrated analysis of fmri and snps. *NeuroImage*, 102:220–228, 2014.

[5] Zaixiang Tang, Yueping Shen, Xinyan Zhang, and Nengjun Yi. The spike-and-slab lasso generalized linear models for prediction and associated genes detection. *Genetics*, 2016.

[6] Thomas Villmann, Frank-Michael Schleif, Markus Kostrzewa, Axel Walch, and Barbara

Hammer. Classification of mass-spectrometric data in clinical proteomics using learning vector quantization methods. *Briefings in Bioinformatics*, 9(2):129–143, 2008.

[1] L. Yeo, N. Adlard, M. Biehl, M. Juarez, T. Smallie, M. Snow, C.D. Buckley, K. Raza, A. Filer, and D. Scheel-Toellner. Expression of chemokines CXCL4 and CXCL7 by synovial macrophages defines an early stage of rheumatoid athritis. *Annals of the Rheumatic Diseases*, 2015. available on-line.

[8] S. Sathiya Keerthi. Generalized LARS as an effective feature selection tool for text classification with svms. In Luc De Raedt and Stefan Wrobel, editors, *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*, volume 119 of *ACM International Conference Proceeding Series*, pages 417–424. ACM, 2005.

[9] Petra Schneider, Michael Biehl, and Barbara Hammer. Adaptive relevance matrices in learning vector quantization. *Neural Computation*, 21(12):3532–3561, 2009.

[10] Hui Zou. An improved 1-norm SVM for simultaneous classification and variable selection. In Marina Meila and Xiaotong Shen, editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, AISTATS 2007, San Juan, Puerto Rico, March 21-24, 2007*, volume 2 of *JMLR Proceedings*, pages 675–681. JMLR.org, 2007.

[11] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artif. Intell.*, 97(1-2):273–324, December 1997.

[12] Benoit Frenay, Daniela Hofmann, Alexander Schulz, Michael Biehl, and Barbara Hammer. Valid interpretation of feature relevance for linear data mappings. In *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 149 – 156. Institute of Electrical & Electronics Engineers (IEEE), 2014.

[13] Alexander Schulz, Bassam Mokbel, Michael Biehl, and Barbara Hammer. Inferring Feature Relevances From Metric Learning. In *2015 IEEE Symposium Series on Computational Intelligence*. Institute of Electrical & Electronics Engineers (IEEE), 2015.

[14] Roland Nilsson, José M. Peña, Johan Björkegren, and Jesper Tegnér. Consistent feature selection for pattern recognition in polynomial time. *J. Mach. Learn. Res.*, 8:589–612, December 2007.

[15] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[16] Miron B. Kursa and Witold R. Rudnicki. The all relevant feature selection using random forest. *CoRR*, abs/1106.5112, 2011.

[17] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1-3):389–422, January 2002.

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

[19] W. H. Wolberg and O. L. Mangasarian. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences*, 87(23):9193–9196, Dec 1990.