A New Approach to Classification with the Least Number of Features

Sascha Klement and Thomas Martinetz
Institute for Neuro- and Bioinformatics, University of Lübeck, Germany
{*klement, martinetz*}*@inb.uni-luebeck.de*

***Abstract*—Recently, the so-called Support Feature Machine (SFM) was proposed as a novel approach to feature selection for classification, based on minimisation of the zero norm of a separating hyperplane. We propose an extension for linearly non-separable datasets that allows a direct trade-off between the number of misclassified data points and the number of dimensions. Results on toy examples as well as real-world datasets demonstrate that this method is able to identify relevant features very effectively.**

***Keywords*-Support feature machine, feature selection, zero norm minimisation, classification.**

I. Introduction

The ever increasing complexity of real-world machine learning tasks requires more and more sophisticated methods to deal with datasets that contain only very few relevant features but many irrelevant noise dimensions. In practise, these scenarios often arise in the analysis of biological datasets, such as tissue classification using microarrays [1], identification of disease-specific genome mutations or distinction between mental states using functional magnetic resonance imaging [2]. It is well-known that a large number of irrelevant features may distract state-of-the-art methods, such as the support vector machine. Thus, feature selection is often a fundamental preprocessing step to achieve proper classification results, to improve runtime, and to make the training results more interpretable.

For many machine learning tasks, maximum margin methods have been confirmed to be a good choice to maximise the generalisation performance [3]. But, besides generalisation capabilities, other aspects, such as fast convergence, existence of simple error bounds, straightforward implementation, running time requirements, or numerical stability, may be equally important.

In recent years, as complexity and dimensionality of real-world problems have dramatically increased, two other aspects have gained more and more importance. These are sparsity and domain interpretability of the inference model. Both are closely connected to the task of variable or feature selection. Primarily, feature selection aims to improve or at least preserve the discriminative capabilities when using fewer features than the original classifier, regression or density estimator. In the following, we focus on feature selection for classification tasks.

Feature selection as an exhaustive search problem is in general computationally intractable as the number of states in the search space increases exponentially with the number of features. Therefore, all computationally feasible feature selection techniques try to approximate the optimal feature set, e.g. by Bayesian inference, gradient descent, genetic algorithms, or various numerical optimisation methods.

Commonly, these methods are divided into two classes: filter and wrapper methods. First, filter methods completely separate the feature selection and the classification task [4]. The feature subset is selected in advance, i.e. filtered out from the overall set of features without assessing the actual classifier.

Wrapper methods make use of the induction algorithm to assess the prediction accuracy of a particular feature subset. Well-known contributions to this class of feature selection algorithms are those of Weston et al. [5], who select those features that minimise bounds on the leave-one-out error, and Guyon et al. [6], who propose the so-called recursive feature elimination. Some types of support vector machines already comprise feature selection to some extend, such as the l_1-norm SVM [7] or the VS-SSVM (Variable Selection via Sparse SVMs) [8]. The influence of an exponentially increasing number of irrelevant features on feature selection in general has been discussed in [9].

Recently, we proposed the so-called Support Feature Machine (SFM) [10] as a novel method to feature selection that is both simple and fast. However, the standard formulation is limited to linearly separable datasets.

The following sections are organised as follows. First, we introduce the problem of finding relevant variables by means of zero norm minimisation. We outline the mathematical formulation of the Support Feature Machine and its extension to linearly non-separable datasets. Additionally, we provide an estimate of incidental separability to answer the question whether the feature selection reveals the fundamental structure of a particular data set or if the same outcome could be observed on random data. An evaluation of the SFM in scenarios with an exponentially increasing number of features and on linearly non-separable datasets follows. Finally, we demonstrate the performance of the SFM on a real-world microarray dataset. We conclude with a critical discussion of the achievements and propose further extensions to the SFM.

II. Feature Selection by Zero-Norm Minimisation

We make use of the common notations used in classification and feature selection frameworks, i.e. the training set

$$\mathcal{D} = \{\vec{x}_i, y_i\}_{i=1}^n$$

consists of feature vectors $\vec{x}_i \in \mathbb{R}^d$ and corresponding class labels $y_i \in \{-1, +1\}$. First, we assume the dataset $\mathcal{D}$ to be linearly separable, i.e.

$$\begin{aligned} & \exists\, \vec{w} \in \mathbb{R}^d,\, b \in \mathbb{R} \\ \text{with} \quad & y_i \left(\vec{w}^{\mathrm{T}} \vec{x}_i + b\right) \geq 0 \;\; \forall\, i \quad \text{and} \quad \vec{w} \neq \vec{0}\,, \end{aligned} \tag{1}$$

where the normal vector $\vec{w} \in \mathbb{R}^d$ and the bias $b \in \mathbb{R}$ describe the separating hyperplane except for a constant factor. Obviously, if $\vec{w}$ and b are solutions to the inequalities, also $\lambda\,\vec{w}$ and $\lambda\,b$ solve them with $\lambda \in \mathbb{R}^+$.

In general, there is no unique solution to (1). Our goal is to find a weight vector $\vec{w}$ and a bias b which solve

$$\begin{aligned} \text{minimise} \quad & \|\vec{w}\|_0^0 \\ \text{subject to} \quad & y_i \left(\vec{w}^{\mathrm{T}} \vec{x}_i + b\right) \geq 0 \quad \text{and} \quad \vec{w} \neq \vec{0} \end{aligned} \tag{2}$$

with $\|\vec{w}\|_0^0 = \mathrm{card}\,\{w_i | w_i \neq 0\}$. Hence, solutions to (2) solve the classification problem (1) using the least number of features. Note, that any solution can be multiplied by a positive factor and is still a solution.

Some attempts have been made to approximate the above problem with a variant of the Support Vector Machine (SVM), e.g. by Weston et al. [11] who

$$\begin{aligned} \text{minimise} \quad & \sum_{j=1}^{d} \ln\left(\epsilon + |w_j|\right) \\ \text{subject to} \quad & y_i \left(\vec{w}^{\mathrm{T}} \vec{x}_i + b\right) \geq 1\,. \end{aligned} \tag{3}$$

with $0 < \epsilon \ll 1$. A local minimum of (3) is found using an iterative scheme based on linear programming. However, we found the following approach to identify relevant features more effectively.

A. Standard Support Feature Machine

Instead of modifying the SVM setting as in [11], we slightly change (2) such that we

$$\begin{aligned} \text{minimise} \quad & \|\vec{w}\|_0^0 \\ \text{subject to} \quad & y_i \left(\vec{w}^{\mathrm{T}} \vec{x}_i + b\right) \geq 0 \\ \text{and} \quad & \vec{w}^{\mathrm{T}} \vec{u} + \bar{y} b = 1 \end{aligned} \tag{4}$$

with $\vec{u} = \frac{1}{n} \sum_{i=1}^n y_i \vec{x}_i$ and $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$. The second constraint excludes $\vec{w} = \vec{0}$, since otherwise we would obtain $\bar{y} b = 1$ and $y_i b \geq 0$, which cannot be fulfilled for all i (we have labels $+1$ and -1). As long as there is a solution to (2) with $y_i \left(\vec{w}^{\mathrm{T}} \vec{x}_i + b\right) > 0$ for at least one $i \in \{1, ..., n\}$, also $\sum_{i=1}^n y_i \left(\vec{w}^{\mathrm{T}} \vec{x}_i + b\right) > 0$ is satisfied. Hence, solving (4) yields a solution to the ultimate problem (2).

Since we have linear constraints, for solving (4) we can employ the same framework Weston et al. [11] used for solving their problem. However, our experiments show that by

$$\begin{aligned} \text{minimising} \quad & \sum_{j=1}^{d} \ln\left(\epsilon + |w_j|\right) \\ \text{subject to} \quad & y_i \left(\vec{w}^{\mathrm{T}} \vec{x}_i + b\right) \geq 0 \quad \text{and} \quad \vec{w}^{\mathrm{T}} \vec{u} + \bar{y} b = 1 \end{aligned}$$

we obtain significantly better solutions to the ultimate problem then by solving (3). It seems that the new cost function is much less prone to local minima. For solving the above problem, we apply a constrained gradient descent technique based on Frank and Wolfe's method [12]:

1) Set $\vec{z} = (1, \ldots, 1)$.
2) Minimise $|\vec{w}|$ such that $y_i \left(\vec{w}^{\mathrm{T}} (\vec{x}_i * \vec{z}) + b\right) \geq 0$ and $\frac{1}{n} \sum_{i=1}^n y_i \left(\vec{w}^{\mathrm{T}} (\vec{x}_i * \vec{z}) + b\right) = 1$
3) Set $\vec{z} = \vec{z} * \vec{w}$.
4) Repeat until convergence.

B. Extension to non-separable Datasets

We extend the SFM to non-separable datasets by introducing a slack variable ξ_i for each data point and a softness parameter C. Then we

$$\begin{aligned} \text{minimise} \quad & \|\vec{w}\|_0^0 + C \|\vec{\xi}\|_0^0 \\ \text{subject to} \quad & \begin{cases} y_i \left(\vec{w}^{\mathrm{T}} \vec{x}_i + b\right) \geq -\xi_i \\ \vec{w}^{\mathrm{T}} \vec{u} + \bar{y} b = \pm 1 \\ \xi_i \geq 0\,. \end{cases} \end{aligned}$$

As we allow for classification errors, $y_i \left(\vec{w}^{\mathrm{T}} \vec{x}_i + b\right)$ may become negative and the pathological case where $\vec{w}^{\mathrm{T}} \vec{u} + \bar{y} b$ gets negative may occur. Therefore, the optimiser needs to fulfil the latter constraint either with $+1$ or -1. Practically, one needs to optimise for both variants and finally choose the solution with the lower objective function. Again, we use the previously mentioned iterative approximation scheme for solving (5).

An appealing feature of the soft-margin SFM is that the objective function explicitly contains the trade-off between the number of features and the number of misclassified training samples $\|\vec{\xi}\|_0^0$.

C. Notes on Incidental Separability

Finally, we want to assess the issue of incidental separability, i.e. the probability of a random dataset to be linearly separable depending on the number of features and the number of data points. In general, there exists no closed formulation for this probability, but in case of rotationally symmetric distributions some bounds can be derived. Let $P_{D,n}$ denote the probability of n data points drawn from a D-dimensional distribution to be linearly separable. This is equivalent to the probability that all data points are

located within the same half-space. Obviously, $P_{D,n} = 1$ for $n \leq D$. For rotationally symmetric distributions, such as the multidimensional standard normal distribution, Wendel [13] proofed that

$$P_{D,n} = 2^{-n+1} \sum_{k=0}^{D-1} \binom{n-1}{k}. \quad (5)$$

Assume a feature selection algorithm indicates that only $d < D$ dimensions are relevant. Now, what is the probability $P_{d,D,n}$ that a d-dimensional subspace exists where all data points are linearly separable or, in other terms, located in the same half-space. As there are $\binom{D}{d}$ possible ways to choose the d-dimensional subspace, the following upper bound holds:

$$\begin{aligned} P_{d,D,n} &\leq \min\left(1, \binom{D}{d} P_{d,n}\right) \\ &\leq \min\left(1, \binom{D}{d} 2^{-n+1} \sum_{k=0}^{d-1} \binom{n-1}{k}\right) \quad (6) \end{aligned}$$

Admittedly, this is a very rough estimate constrained to the strong requirement of rotationally symmetric distributions. However, if $P_{d,D,n}$ is low in an arbitrary scenario, it is a strong indicator that the selected features are truly relevant. In other words, it is not likely that a random data set with the same parameters is separable by chance.

Finally, we want to address the special case for $d = 1$ and the multidimensional standard normal distribution. Let E_i denote the event that the dataset is separable within dimension i. Now, the probability $P_{1,D,n}$ derives to

$$\begin{aligned} P_{1,D,n} &= P\left(\bigcup_{i=1}^{D} E_i\right) \\ &= P(E_1) + \ldots + P(E_D) \\ &\quad -P(E_1 \cap E_2) - \ldots - P(E_{D-1} \cap E_D) \\ &\quad +P(E_1 \cap E_2 \cap E_3) + \ldots \\ &\quad \ldots \\ &\quad (-1)^{D-1} P\left(\bigcap_{i=1}^{D} E_i\right) \\ &= \sum_{i=1}^{D} (-1)^{i+1} \binom{D}{i} P_{1,n}^{i} \\ &= \sum_{i=1}^{D} (-1)^{i+1} \binom{D}{i} 2^{i\cdot(-n+1)} \quad (7) \end{aligned}$$

Here, we use the fact that all events E_i are pairwise statistically independent, i.e. $P(E_i \cap E_j) = P(E_i)P(E_j)$ for all $i \neq j$.

III. Experiments

For learning tasks, such as classification or regression, one normally assesses a method's performance via the k-fold cross-validation error, or via the test error on a separate dataset. For feature selection, besides the test error, also the number of selected features and the amount of truly relevant features are important. As in real-world scenarios these values are almost never known, we start with artificial examples to compare the results of different methods.

A. Exponentially Increasing Number of Irrelevant Features

First, we focus on the impact of an exponentially increasing number of irrelevant features as this is the most interesting scenario in real-world machine learning and pattern recognition applications, such as the analysis of microarray or genome data. We normally deal with an extremely large number of input features while the number of data points is low due to practical and financial issues in data acquisition. However, we expect the number of relevant dimensions in these scenarios to be rather low with respect to the whole number of input dimensions.

The toy examples were constructed according to Weston et al. [11], i.e. the input data consist of 6 relevant but redundant features and an exponentially increasing number of Gaussian noise dimensions k^* (k^* = 8, 64, 512, 4096, the original dataset contained a fixed number of 196 noise dimensions). We sampled 10000 data points, a small proportion of n data points was used for training ($n = 20, 50, 100, 200, 500$), the remaining data points served as the test set. Additionally, we required the training set to be linearly separable. Each experiment was conducted 100 times for each training method.

Table I shows the impact of the noise features on the test error for the SVM without feature selection, Weston's method and the SFM. Obviously, the SVM shows a very bad performance for $k^* \geq 512$. Both, the SFM and Weston's method, are significantly better suited in these scenarios and Weston's method shows the lowest error rate in extremely low-dimensional scenarios (e.g. $n = 20, k^* = 512$). However, the test error of the SFM increases only slowly with the number of irrelevant features k^* and the increase from $k^* = 512$ to $k^* = 4096$ is well below the standard deviation.

Table II compares the capability of both variable selection methods to identify relevant and irrelevant features. Obviously, the SFM returns a lower number of features which are more likely to be truly relevant features. Even in high-dimensional low-sample size scenarios the SFM can identify very effectively the relevant dimensions. As the number of data points increases, the number of features we find to be relevant increases but stays below 6 — the number of truly relevant features. The percentage of correctly identified features decreases with the number of noise dimension. However, only in extremely high-dimensional low-sample size scenarios the value drops below 90%.

Finally we apply the estimate for incidental separability (6). For $n = 20$ and $k^* = 4096$, we find $2.2 \approx 2$ relevant dimensions. Here, $P_{d,D,n} = P_{2,5102,20} = 1$, i.e. the result does not reveal significant structure — empirically

only 40.9% of the identified features were truly relevant on average. However, for $n = 50$ and again $k^* = 4096$, approximately 3 features were found. Now, $P_{3,5102,50} = 0.048$, being a strong indicator that these features indeed reveal structure. Empirically, 84.7% were correctly identified on average.

B. Soft-Margin Support Feature Machine

For evaluating the soft-margin approach, we constructed an artificial problem where both classes have a significant overlap. The probability of the classes $y = 1$ and $y = -1$ was equal both in the training and the test set. The first k dimensions $x_1, \ldots, x_k$ were drawn normally distributed as $x_i = \mathcal{N}(\mu \cdot y, 1)$. The remaining features $x_{k+1}, \ldots, x_d$ were noise drawn as $x_i = \mathcal{N}(0, 1)$. Training and test sets were sampled according to the above procedure, each containing n data points.

Figure 1 shows the mean results after 100 repetitions for $n = 500, k = 10, d = 200$ and $\mu = 0.3$. The softness parameter C was sampled in 100 steps logarithmically spaced between 0.01 and 100.

We observe the number of features to increase with C. For very small C exclusively relevant features are selected, hence, a correct feature rate of 100% is achieved. As the number of features approaches the true number of features ($k = 10$) more and more irrelevant features are included. The training error decreases with C but does not become zero as both classes have a significant overlap. The test error is minimal for $C = 0.66$.

C. Microarray Data

In recent years, the use of microarray data has become an important tool to determine genes that cause certain diseases. Golub et al. [1] showed how to classify two different types of cancer (acute myeloid leukaemia (AML) and acute lymphoblastic leukaemia (ALL)) using correlation coefficients. The dataset consists of 38 training samples (27 vs. 11) and 34 test samples (20 vs. 14) with 7129 features each describing the expression level of a single protein.

An important aspect of this high-dimensional low-sample size datasets is the strong correlation between some of the relevant features. Using an SFM, we find the training dataset to be separable in two dimensions, but with a test error of 17.7%. This is due to two different aspects. First, the SFM may have multiple solutions, so there may exist other two dimensional projections that are also linearly separable. Second, whenever dimensions are strongly correlated, the SFM selects one of them if separability is achieved. However, it might be beneficial with respect to generalisation performance to also include these correlated features. We propose the following greedy method for identifying correlated dimensions:

1) Initialise the set of active features $\mathcal{A} = \{1, \ldots, d\}$ and the set of relevant features $\mathcal{F} = \emptyset$.

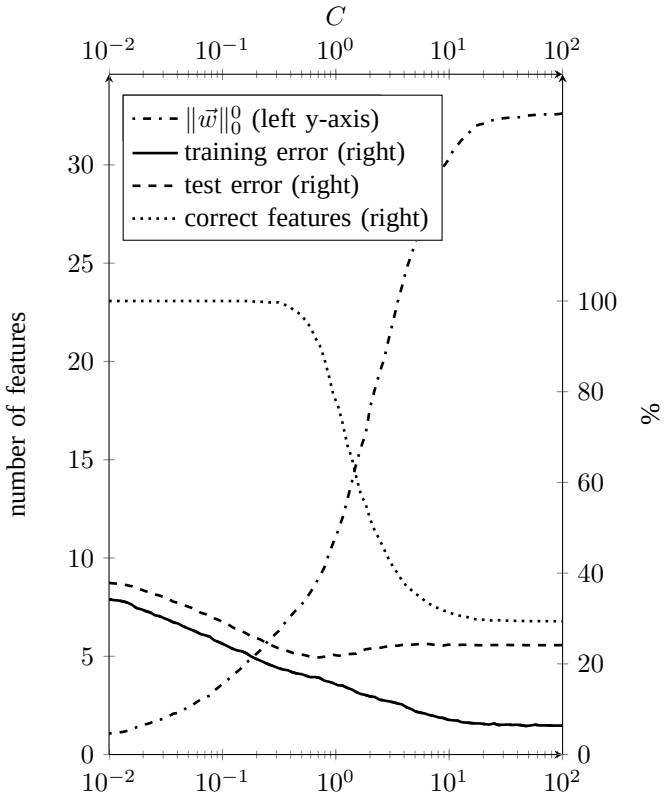

Figure 1. Soft margin SFM ($n = 500, k = 10, d = 200$). Shown are the number of features $\|\vec{w}\|_0^0$ (dash-dotted, left y-axis), the percentage of misclassified training data points (solid, right y-axis), the test error (dashed, right y-axis) and the percentage of correctly identified features (dotted, right y-axis) for 100 independent trials.

2) Train an SFM using the feature set $\mathcal{A}$.
3) Remove all features with non-zero weight from $\mathcal{A}$ and add these features to $\mathcal{F}$.
4) Repeat for a fixed number of iterations or until the training error exceeds a certain threshold.

The final feature set $\mathcal{F}$ may now be used for training an SVM to optimise for generalisation performance. After 10 iterations ($C = 1000$) we obtained a total of 25 features. A soft-margin SVM ($C = 10000$) using these 25 features misclassified only a single test data point.

We further compared the obtained list of relevant features with the 50 most relevant features in [1]. We found that 20 of the 25 features were also present in the feature list of Golub et al. As both methods rely on different theoretical approaches — Golub et al. use a ranking of correlation coefficients, we use a multidimensional optimisation procedure — the results are not only consistent but also emphasise the biological relevance of the selected feature set. These are promising results for future applications of the SFM.

Table I

IMPACT OF AN EXPONENTIALLY INCREASING NUMBER OF IRRELEVANT FEATURES ON THE TEST ERROR. SHOWN ARE THE TEST ERROR OF AN SVM WITHOUT FEATURE SELECTION (A), OF THE METHOD PROPOSED BY WESTON (B) AND OF THE SFM (C) (TOY EXAMPLE, 6 RELEVANT FEATURES, $k*$ IRRELEVANT ONES).

(a) Support Vector Machine

n	$k^* = 8$	$k^* = 64$	$k^* = 512$	$k^* = 4096$
20	5.0% (± 2.5%)	17.5% (± 2.8%)	36.1% (± 1.3%)	45.1% (± 0.7%)
50	2.4% (± 1.2%)	8.7% (± 1.7%)	28.6% (± 1.3%)	41.9% (± 0.7%)
100	1.7% (± 0.8%)	5.0% (± 1.3%)	21.6% (± 1.3%)	38.7% (± 0.7%)
200	1.3% (± 0.5%)	2.7% (± 0.7%)	14.1% (± 1.0%)	34.2% (± 0.7%)
500	1.0% (± 0.3%)	1.6% (± 0.3%)	6.9% (± 0.7%)	26.2% (± 0.6%)

(b) Weston's Method

n	$k^* = 8$	$k^* = 64$	$k^* = 512$	$k^* = 4096$
20	6.0% (± 5.7%)	6.8% (± 6.8%)	14.5% (±12.0%)	30.9% (±15.6%)
50	2.3% (± 1.8%)	2.6% (± 1.7%)	3.0% (± 2.2%)	3.7% (± 3.0%)
100	1.6% (± 0.7%)	1.8% (± 1.0%)	1.8% (± 0.9%)	1.8% (± 1.0%)
200	1.2% (± 0.5%)	1.2% (± 0.5%)	1.3% (± 0.5%)	1.3% (± 0.5%)
500	0.8% (± 0.3%)	0.8% (± 0.2%)	0.8% (± 0.3%)	0.8% (± 0.3%)

(c) Support Feature Machine

n	$k^* = 8$	$k^* = 64$	$k^* = 512$	$k^* = 4096$
20	14.5% (± 5.5%)	14.7% (± 5.9%)	23.0% (±11.8%)	31.5% (±14.9%)
50	5.8% (± 2.7%)	6.4% (± 3.9%)	7.2% (± 4.1%)	8.5% (± 4.8%)
100	3.4% (± 1.7%)	3.2% (± 1.5%)	3.3% (± 1.5%)	3.6% (± 1.8%)
200	1.9% (± 0.8%)	2.0% (± 0.7%)	2.1% (± 0.8%)	1.9% (± 0.8%)
500	1.1% (± 0.3%)	1.1% (± 0.4%)	1.1% (± 0.4%)	1.0% (± 0.4%)

Table II

IMPACT OF AN EXPONENTIALLY INCREASING NUMBER OF IRRELEVANT FEATURES ON THE VARIABLE SELECTION PERFORMANCE. SHOWN ARE THE NUMBER OF FEATURES FOUND TO BE RELEVANT (A,B) AND THE PERCENTAGE OF CORRECTLY IDENTIFIED FEATURES (C,D).

(a) SFM, features found to be relevant

n	$k^* = 8$	$k^* = 64$	$k^* = 512$	$k^* = 4096$
20	2.0 (± 0.6)	2.1 (± 0.7)	2.1 (± 0.7)	2.2 (± 0.8)
50	2.4 (± 0.6)	2.5 (± 0.6)	2.5 (± 0.8)	2.6 (± 0.8)
100	2.6 (± 0.7)	2.7 (± 0.7)	2.7 (± 0.7)	2.7 (± 0.7)
200	3.1 (± 0.7)	3.3 (± 0.8)	3.0 (± 0.8)	3.1 (± 0.8)
500	4.2 (± 0.8)	4.1 (± 0.7)	4.0 (± 0.7)	4.0 (± 0.9)

(b) Weston, features found to be relevant

n	$k^* = 8$	$k^* = 64$	$k^* = 512$	$k^* = 4096$
20	2.8 (± 0.8)	2.7 (± 0.9)	3.1 (± 1.1)	3.4 (± 1.3)
50	3.2 (± 1.0)	3.2 (± 1.1)	3.4 (± 1.2)	3.5 (± 1.4)
100	3.8 (± 1.0)	4.0 (± 1.2)	3.8 (± 1.3)	3.8 (± 1.2)
200	4.8 (± 1.3)	4.8 (± 1.3)	5.0 (± 1.2)	4.8 (± 1.4)
500	5.8 (± 1.3)	6.2 (± 1.5)	6.0 (± 1.4)	6.0 (± 1.3)

(c) SFM, Correctly identified relevant features

n	$k^* = 8$	$k^* = 64$	$k^* = 512$	$k^* = 4096$
20	98.0% (± 9.0%)	85.6% (±22.8%)	67.2% (±26.5%)	40.9% (±34.4%)
50	98.6% (± 6.1%)	99.4% (± 4.1%)	94.4% (±14.0%)	84.7% (±23.6%)
100	99.8% (± 2.5%)	99.5% (± 3.5%)	97.1% (± 9.4%)	94.9% (±12.0%)
200	99.6% (± 2.6%)	98.8% (± 5.1%)	97.4% (± 7.5%)	96.3% (± 9.5%)
500	98.6% (± 5.2%)	96.6% (± 8.0%)	95.6% (± 8.6%)	94.3% (±10.5%)

(d) Weston, Correctly identified relevant features

n	$k^* = 8$	$k^* = 64$	$k^* = 512$	$k^* = 4096$
20	93.0% (±15.1%)	82.8% (±20.4%)	62.7% (±29.9%)	33.9% (±30.1%)
50	93.8% (±12.7%)	90.2% (±14.6%)	81.0% (±18.7%)	82.3% (±20.1%)
100	95.0% (±10.0%)	91.1% (±14.6%)	88.0% (±16.8%)	85.3% (±15.8%)
200	90.3% (±12.8%)	82.2% (±15.0%)	79.6% (±16.3%)	82.7% (±15.9%)
500	83.8% (±12.9%)	76.9% (±15.0%)	75.4% (±14.8%)	78.6% (±15.2%)

D. Implementation Issues

As with many machine learning algorithms, normalisation is an essential preprocessing step also for the SFM. For all experiments, we normalised the training datasets to zero mean and unit variance and finally scaled all vectors to have a mean norm of one. This last step is necessary in high-dimensional scenarios to keep the outcome of scalar products in a reasonable range. The test sets were normalised according to the factors obtained from the corresponding training sets.

For solving the optimisation problems, we used the MOSEK optimisation software. To avoid numerical issues, numbers that differed by no more than a specific implementation-dependent number — normally closely connected to the machine epsilon — were considered to be equal.

In the hard-margin case, either no solution exists or a solution where all data points are correctly classified. Since the optimiser uses numerical approximation methods with certain accuracy thresholds, some constraints may be marginally violated. Thus, some data points may be located on the wrong side of the hyperplane, but very close to it, producing a non-zero training error even in the hard-margin case.

IV. Conclusions

Experiments on artificial as well as real-world datasets demonstrated that the SFM can identify relevant features very effectively and may improve the generalisation performance significantly with respect to an SVM without feature selection. Even an exponentially increasing number of irrelevant features does not cause a significant performance drop. The implementation only requires linear programming solvers and may therefore be established in various programming environments.

Additionally, we introduced some simple bounds for the probability of incidental separability, that may be used to estimate whether separability in a certain scenario is likely or may occur even for a completely random data set.

So far, we focused on linear classifiers, mostly for high-dimensional low-sample size scenarios because these scenarios seem to be the most relevant ones in practical applications, such as the analysis of microarray datasets.

In some scenarios, it is necessary to allow for nonlinear classification to achieve proper classification performance. One might think of ways to incorporate kernels into the SFM to allow for arbitrary class boundaries. Nevertheless, the main focus of the SFM was to provide results that may easily be interpreted both in terms of feature selection and classification, so nonlinearities would slacken this demand.

Further work will include experiments on more challenging real-world scenarios with practical relevance. Finally, we seek for an iterative optimisation method to be independent from proprietary optimisation toolboxes.

References

[1] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, and C. D. Bloomfield, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531–537, 1999.

[2] J.-D. Haynes and G. Rees, "Decoding mental states from brain activity in humans," *Nature Reviews Neuroscience*, vol. 7, pp. 523–534, 2006.

[3] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer, 1995.

[4] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, pp. 273–323, 1997.

[5] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature Selection for SVMs," in *Advances in Neural Information Processing Systems*, 2000.

[6] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389–422, 2002.

[7] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, "1-norm support vector machines," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2004.

[8] J. Bi, K. P. Bennett, M. Embrechts, C. M. Breneman, and M. Song, "Dimensionality Reduction via Sparse Support Vector Machines," *Journal of Machine Learning Research*, vol. 3, pp. 1229–1243, 2003.

[9] A. Y. Ng, "On feature selection: learning with exponentially many irrelevant features as training examples," in *Proc. 15th International Conf. on Machine Learning*. Morgan Kaufmann, San Francisco, CA, 1998, pp. 404–412. [Online]. Available: citeseer.ist.psu.edu/ng98feature.html

[10] S. Klement and T. Martinetz, "The support feature machine for classifying with the least number of features," in *ICANN (2)*, ser. Lecture Notes in Computer Science, K. I. Diamantaras, W. Duch, and L. S. Iliadis, Eds., vol. 6353. Springer, 2010, pp. 88–93.

[11] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping, "Use of the Zero-Norm with Linear Models and Kernel Methods," *Journal of Machine Learning Research*, vol. 3, pp. 1439–1461, 2003.

[12] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Naval Research Logistics Quarterly*, vol. 3, pp. 95–110, 1956.

[13] J. Wendel, "A problem in geometric probability," *Mathematics Scandinavia*, vol. 11, pp. 109–111, 1962.