# A Short Note About the Application of Polynomial Kernels with Fractional Degree in Support Vector Learning 

Rolf Rossius, Gérard Zenker, Andreas Ittner, and Werner Dilger<br>Department of Computer Science<br>Artificial Intelligence Group<br>Chemnitz University of Technology<br>D-09107 Chemnitz<br>\{ros,gze,ait,wdi\}@informatik.tu-chemnitz.de<br>http://www.tu-chemnitz.de/informatik/HomePages/KI/


#### Abstract

In the mid 90's a fundamental new Machine Learning approach was developed by V. N. Vapnik: The Support Vector Machine (SVM). This new method can be regarded as a very promising approach and is getting more and more attention in the fields where neural networks and decision tree methods are applied. Whilst neural networks may be considered (correctly or not) to be well understood and are in wide use, Support Vector Learning has some rough edges in theoretical details and its inherent numerical tasks prevent it from being easily applied in practice. This paper picks up a new aspect - the use of fractional degrees on polynomial kernels in the SVM - discovered in the course of an implementation of the algorithm. Fractional degrees on polynomial kernels broaden the capabilities of the SVM and offer the possibility to deal with feature spaces of infinite dimension. We introduce a method to simplify the quadratic programming problem, as the core of the SVM.


## 1 Introduction

Well known representatives of classification and prediction methods in the field of Machine Learning are neural networks and methods for generation different kinds of decision trees. An innovative and still relatively unknown learning approach is the Support Vector Machine (SVM) developed by V. N. Vapnik in the mid 90 's. Support Vector Learning [IRZ98] is not just another approach to learning techniques, rather it can be regarded as a fundamental new philosophy in the area of Machine Learning.

The underlying principle of the SVM is the principle of the Structural Risk Minimization (SRM) [Vap95]. In contrast to a pure minimization of the empirical risk the SRM is based on the "idea of the simplicity" and unifies Empirical Risk Minimization and the problem of Model Selection. The searched binary classifier for the problem

$$
\left(x_{1}, y_{1}\right), \ldots,\left(x_{l}, y_{l}\right), x_{i} \in \mathbf{R}^{n}, y_{i} \in\{+1,-1\}
$$

has to be a function from the set

$$
\left\{f_{\alpha}: \alpha \in \Gamma\right\}, f_{\alpha}: \mathbf{R}^{n} \rightarrow\{+1,-1\}, x \mapsto y
$$

and should reflect the real inherent essence of the given learning problem. This essence can be regarded as the simplest (in some sense) separation of the feature space. Here simplicity will be formalized by means of the $V C$ dimension, i. e. a measure of the considered set of feasible functions, e. g. the family of separating hyperplanes. The SRM is enforced by controlled bounding of the VC dimensions of the set $\left\{f_{\alpha}\right\}$ and ensures the excellent generalization ability of the SVM. The underlying theory of the SRM will not be explained in detail in this paper. We refer to [Vap95] which covers the SRM and the application in the SVM.

The separating hyperplane is characterized by $\langle\omega, x\rangle+b=0$. The distance between the hyperplane and the examples should be maximized, i. e. one has to solve a problem of mathematical programming. For the non-separable case slack variables $\xi_{i} \geq 0$ are introduced, which leads to:

$$
\begin{cases}\frac{1}{2}\langle\omega, \omega\rangle+C \sum_{i=1}^{l} \xi_{i} \rightarrow \min &  \tag{1}\\ y_{i}\left[\left\langle\omega, x_{i}\right\rangle+b\right] \geq 1-\xi_{i} & \forall i=1, \ldots, l \\ \xi_{i} \geq 0 & \forall i=1, \ldots, l\end{cases}
$$

where the capacity parameter $C>0$ controls the interrelationship between the accuracy of the classifier on the learning set and its ability of generalization, i. e. the accuracy on an unseen test set.

The vector $\omega$, as the solution of (1), determines the optimal hyperplane. It can be expressed as a linear combination of a possibly small subset of the whole learning data:

$$
\begin{equation*}
\omega=\sum_{i=1}^{l} \alpha_{i} y_{i} x_{i}=\sum_{S V} \alpha_{i} y_{i} x_{i} \tag{2}
\end{equation*}
$$

Support Vectors are such vectors $x_{i}$, which satisfy $y_{i}\left[\left\langle\omega, x_{i}\right\rangle+b\right]=1$, i. e. which have a nonzero $\alpha_{i}$ and effectively contribute to the description of the separating hyperplane. Hence in (2) one can reduce $\omega$ to a linear combination of support vectors. Less formally these support vectors can be viewed as the examples on the frontline guarding the own class against the examples of the other one and are essential for the concept to be learned.

Considering (2) one has to solve the following optimization problem:

$$
\left\{\begin{array}{l}
\Lambda^{T} 1-\frac{1}{2} \Lambda^{T} A \Lambda \rightarrow \max  \tag{3}\\
0 \leq \Lambda \leq C 1 \\
\Lambda^{T} Y=0
\end{array}\right.
$$

with $\Lambda=\left(\alpha_{1}, \ldots, \alpha_{l}\right), \mathbf{1}=(1, \ldots, 1)$, and $Y=\left(y_{1}, \ldots, y_{l}\right)$. The Hesse matrix $A$ consists of the elements $A_{i j}=y_{i} y_{j}\left\langle x_{i}, x_{j}\right\rangle$ for $i, j=1, \ldots, l$ [CV95].

However in the general case the linear separation in the original feature space will not provide a sufficient classifier. Therefore the original feature space is
expanded to a very high dimensional image space by (e.g.):
$\Phi: \mathbf{R}^{n} \rightarrow \mathbf{R}^{N}, n \ll N, \Phi(x)=\left(1, \gamma_{1} x_{1}, \ldots, \gamma_{n} x_{n}, \gamma_{n+1} x_{1}^{2}, \gamma_{n+2} x_{1} x_{2}, \ldots, \gamma_{k} x_{n}^{d}\right)$,
and in this space the linear separation is performed. An inverse transformation back into $\mathbf{R}^{n}$ results in a non-linear separation in the original space of the task supplied features:

$$
f(x)=\langle\omega, \Phi(x)\rangle+b
$$

It is not necessary to expand the feature space explicitly. One way to do the mapping implicitly is to use kernels $K(u, v)$ (respectively dot products). In this context the fundamental interrelation is:

$$
K(u, v)=\langle\Phi(u), \Phi(v)\rangle
$$

The symmetric function $K(u, v)$ may be a dot product for the high dimensional image space, if the eigenvalues are positive. One rather simple type of such kernels is representable as

$$
\begin{equation*}
K(u, v)=(\langle u, v\rangle+1)^{d}, \quad d=1,2, \ldots \tag{4}
\end{equation*}
$$

with degree $d$ as an integer. Another choice may be $K(u, v)=e^{-\frac{\|u-u\|}{\sigma}}$. A generalized kind of the kernel (4) will be examined in this paper.

## 2 Polynomial Kernels with Fractional Degree

Interestingly a fixed chosen kernel $K(u, v)$ induces not only exactly one transformation but a manifold of such mappings $\Phi$. Even the dimensionality of the image space $\mathbf{R}^{N}$ is not determined. From (4) for $d=2$ and $n=2$ one gets:

$$
\Phi(u)=\left(1, \sqrt{2} u_{1}, \sqrt{2} u_{2}, u_{1}^{2}, \sqrt{2} u_{1} u_{2}, u_{2}^{2}\right), \quad u=\left(u_{1}, u_{2}\right)
$$

as well as

$$
\Phi(u)=\left(1, u_{1}, u_{1}, u_{2}, u_{2}, u_{1}^{2}, u_{1} u_{2}, u_{2} u_{1}, u_{2}^{2}\right)
$$

and infinite number of others.
Therefore a question arises: Choosing a kernel $K(u, v)$ - which is the space of smallest dimension for an image of $\Phi$ ? The answer for $d \in N$ is $\binom{n+d}{d}$ (or equivalently $\binom{n+d}{n}$ ). While selecting an appropriate kernel $K$ via the exponent $d$, there are huge discontinuities in the dimensionalities of the corresponding image spaces. The approximation and generalization capacity may be controlled by bounding the norm of the separating hyperplane, but another tuning parameter will still be there: the dimensionality (cf. Table 1).

Using a fractional exponent in the kernel (4) we encounter an interesting property: the dot product $\langle u, v\rangle$ may be less than -1 and we have a negative

| $n \backslash d$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 2 | 2 | 6 | 10 | 15 | 21 | 28 | 36 |
| 16 | 16 | 153 | 969 | $4.8 \times 10^{3}$ | $2.0 \times 10^{4}$ | $7.5 \times 10^{4}$ | $2.5 \times 10^{5}$ |
| 256 | 256 | $3.3 \times 10^{4}$ | $2.9 \times 10^{6}$ | $1.9 \times 10^{8}$ | $9.7 \times 10^{9}$ | $4.2 \times 10^{11}$ | $1.6 \times 10^{13}$ |

Table 1. Dimension of image space for polynomial kernel with exponent $d$ and $n$ original features. The dimension of the image space (where the linear separation takes place) grows quite rapidly - an explicit computation in this space would be impossible. But as mentioned before, this is fortunately not required. Rather the value itself should guide the user to a conjecture about the separating abilities of the associated hyperplane.
base to raise. ${ }^{1}$ Hence the Hesse Matrix $A$ will not be real valued and therefore symmetric ( $A^{T}=A$ ) anymore, but in fact contain complex entries. Nevertheless, $A$ has the property of hermiticity $\left(A^{T}=\bar{A}\right)$. This allows for a new formulation of (3). Because

$$
\Lambda^{T} A \Lambda=\Lambda^{T} A^{T} \Lambda=\Lambda^{T} \frac{1}{2}\left(A+A^{T}\right) \Lambda=\Lambda^{T} \frac{1}{2}(A+\bar{A}) \Lambda=\Lambda^{T} \operatorname{Re}(A) \Lambda
$$

we equivalently solve

$$
\left\{\begin{array}{l}
\Lambda^{T} 1-\frac{1}{2} \Lambda^{T} \operatorname{Re}(A) \Lambda \rightarrow \max  \tag{5}\\
0 \leq \Lambda \leq C 1 \\
\Lambda^{T} Y=0
\end{array}\right.
$$

instead, and get rid of the complex entries. $(\operatorname{Re}(A)$ denotes the real part.)
Exposing the kernel for arbitrary exponents $d$ we get according to Taylor:

$$
\begin{aligned}
(\langle u, v\rangle+1)^{d}= & 1+d\langle u, v\rangle+\frac{d(d-1)}{2!}\langle u, v\rangle^{2}+\frac{d(d-1)(d-2)}{3!}\langle u, v\rangle^{3} \\
& +\frac{d(d-1)(d-2)(d-3)}{4!}\langle u, v\rangle^{4}+\ldots
\end{aligned}
$$

Non-integer exponents do not terminate the series like the integer ones, but the influence of high-order terms decreases nevertheless. In contrast to kernels with an integer exponent there are no mappings $\Phi$ corresponding to such a fractional exponent kernel which have an image space of finite dimension.

Fractional degrees allow a more continuous range of concepts. The resulting separating hyperplanes smoothly change the shapes with the exponent $d$. This will be of importance especially for domains dealing with feature spaces which already cover tens, hundreds or more dimensions (e.g. recognition of graphical images), where a lower degree of a polynomial kernel is preferred. A simple artificial problem in a two dimensional feature space is presented in Figure 1. [Fri93]

[^0]

Fig. 1. Continuous variation of exponent d. 226 examples, class distribution 93/133, $90 \%$ used to generate the separation. Two properties of the problem are significant: low dimensionality of the original feature space, difficultly crossed arrangement of examples in the lower right area. As expected, a somewhat higher exponent of the polynomial kernel is necessary for the approximation of the concept.

## 3 The " $1 / 2$ Trick"

Realizing the SVM as a whole, the solution of the quadratic optimization problem (quadratic programming, QP) - actually a series of such, with different parameters - constitutes the real amount of work. Generally the QP task is for the most part determined by the calculation of function values, gradients (or its estimations). It makes more difficulties here because of the (potential) large Hesse matrix and its nonsparsity.

We tackle this by choosing a kernel of the type $(\langle u, v\rangle+1)^{d}$ with $d=m+\frac{1}{2}$ and $m \in \mathbf{N}$. The corresponding entry in the resulting Hesse matrix $(\operatorname{Re}(A)$ in (5)) will vanish for negative $(\langle u, v\rangle+1)$.

The SVM algorithm selects a separating hyperplane according to a criterion of sufficient values on the training examples as well as the minimization of the norm of the hyperplane. Unfortunately, $\Phi$ is nonlinear - the resulting shape of the function and thus the border between the predicted areas of both classes varies with uniform translations of the examples in the feature space. For instance the resulting separation lines for different centered sets of the well known XOR problem are depicted in Figure 2. A second degree kernel is used.

Despite of the non-invariance against the uniform translation of the examples in the feature space, one could center the set into the origin of the co-ordinate system to obtain a sufficient obtuse angle between a large number of pairs of examples. This will result in a sparser Hesse matrix for the QP task. Up to $50 \%$ of the entries may be zeroed by means of this smart approach.


Fig. 2. Noninvariance of separation with respect to translation. $\left(x-\frac{1}{2}, y-\frac{1}{2}\right)$, $\left(x+\frac{1}{2}, y+\frac{1}{2}\right)$ are members of one class, while the two other examples $\left(x-\frac{1}{2}, y+\frac{1}{2}\right)$, $\left(x+\frac{1}{2}, y-\frac{1}{2}\right)$ belong to a second class. The four points are centered on $(x, y)$.

## 4 Summary

The Support Vector algorithm shows some promising properties but needs some refinement especially on the level of practical realization to soften the enormous effort to find the "simplest" explanation for a learning problem. Polynomial kernels with fractional degrees provide a broader range of concepts as well as a way to reduce the numerical effort to be spent in the QP.

The algorithm works well with a feature space of "similar" features. Is is often preferred to do a componentwise transformation to normalize the data in front of the number crunching task of the SVM itself. For specific domains this could be done in the kernel function.

## References

[CV95] C. Cortes and V. N. Vapnik. Support-vector networks. Machine Learning, 20:273-297, 1995.
[Fri93] B. Fritzke. Growing cell structures - a self-organizing network for unsupervised and supervised learning. Technical Report 93-026, International Computer Science Institute, Berkeley, California, 1993.
[IRZ98] A. Ittner, R. Rossius, and G. Zenker. Support Vector Learning. Technical Report CSR-98, Chemnitz University of Technology, Chemnitz, Germany, 1998.
[Vap95] V. N. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, 1995.


[^0]:    ${ }^{1}$ One could imagine this in the original space: The representing vectors $u$ and $v$ of both participating examples form a sufficient obtuse angle.

