A Kernel Approach for Learning from almost Orthogonal Patterns

Bernhard Schölkopf[1], Jason Weston[1], Eleazar Eskin[2], Christina Leslie[2], and William Stafford Noble[2,3]

[1] Max-Planck-Institut für biologische Kybernetik, Spemannstr. 38, D-72076 Tübingen, Germany
{bernhard.schoelkopf, jason.weston}@tuebingen.mpg.de
[2] Department of Computer Science
Columbia University, New York
{eeskin,cleslie, noble}@cs.columbia.edu
[3] Columbia Genome Center
Columbia University, New York

Abstract. In kernel methods, all the information about the training data is contained in the Gram matrix. If this matrix has large diagonal values, which arises for many types of kernels, then kernel methods do not perform well. We propose and test several methods for dealing with this problem by reducing the dynamic range of the matrix while preserving the positive definiteness of the Hessian of the quadratic programming problem that one has to solve when training a Support Vector Machine.

1 Introduction

Support Vector Machines (SVM) and related kernel methods can be considered an approximate implementation of the structural risk minimization principle suggested by Vapnik (1979). To this end, they minimize an objective function containing a trade-off between two goals, that of minimizing the training error, and that of minimizing a regularization term. In SVMs, the latter is a function of the margin of separation between the two classes in a binary pattern recognition problem. This margin is measured in a so-called feature space $\mathcal{H}$ which is a Hilbert space into which the training patterns are mapped by means of a map

$$\Phi : \mathcal{X} \to \mathcal{H}. \tag{1}$$

Here, the input domain $\mathcal{X}$ can be an arbitrary nonempty set. The art of designing an SVM for a task at hand consist of selecting a feature space with the property that dot products between mapped input points, $\langle \Phi(x), \Phi(x') \rangle$, can be computed in terms of a so-called *kernel*

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle \tag{2}$$

T. Elomaa et al. (Eds.): ECML, LNAI 2430, pp. 511-528, 2002.

which can be evaluated efficiently. Such a kernel necessarily belongs to the class of *positive definite kernels* (e.g. Berg et al. (1984)), i.e., it satisfies

$$\sum_{i,j=1}^{m} a_i a_j k(x_i, x_j) \geq 0 \tag{3}$$

for all $a_i \in \mathbb{R}, x_i \in \mathcal{X}, i = 1, \ldots, m$. The kernel can be thought of as a nonlinear similarity measure that corresponds to the dot product in the associated feature space. Using k, we can carry out all algorithms in $\mathcal{H}$ that can be cast in terms of dot products, examples being SVMs and PCA (for an overview, see Schölkopf and Smola (2002)).

To train a hyperplane classifier in the feature space,

$$f(x) = \operatorname{sgn}(\langle \mathbf{w}, \Phi(x) \rangle + b), \tag{4}$$

where $\mathbf{w}$ is expanded in terms of the points $\Phi(x_j)$,

$$\mathbf{w} = \sum_{j=1}^{m} a_j \Phi(x_j), \tag{5}$$

the SVM pattern recognition algorithm minimizes the quadratic form[4]

$$\|\mathbf{w}\|^2 = \sum_{i,j=1}^{m} a_i a_j K_{ij} \tag{6}$$

subject to the constraints

$$y_i \left[\langle \Phi(x_i), \mathbf{w} \rangle + b \right] \geq 1, \text{ i.e., } \quad y_i \left[\sum_{j=1}^{m} a_j K_{ij} + b \right] \geq 1 \tag{7}$$

and

$$y_i a_i \geq 0 \tag{8}$$

for all $i \in \{1, \ldots, m\}$. Here,

$$(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \{\pm 1\} \tag{9}$$

are the training examples, and

$$K_{ij} := k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle \tag{10}$$

is the Gram matrix.

Note that the regularizer (6) equals the squared length of the weight vector $\mathbf{w}$ in $\mathcal{H}$. One can show that $\|\mathbf{w}\|$ is inversely proportional to the margin of

[4] We are considering the zero training error case. Nonzero training errors are incorporated as suggested by Cortes and Vapnik (1995). Cf. also Osuna and Girosi (1999).

separation between the two classes, hence minimizing it amounts to maximizing the margin. Sometimes, a modification of this approach is considered, where the regularizer

$$\sum_{i=1}^{m} a_i^2 \tag{11}$$

is used instead of (6). Whilst this is no longer the squared length of a weight vector in the feature space $\mathcal{H}$, it is instructive to re-interpret it as the squared length in a different feature space, namely in $\mathbb{R}^m$.

To this end, we consider the feature map

$$\Phi_m(x) := (k(x, x_1), \ldots, k(x, x_m))^\top, \tag{12}$$

sometimes called the *empirical kernel map* (Tsuda, 1999; Schölkopf and Smola, 2002). In this case, the SVM optimization problem consists in minimizing

$$\|\mathbf{a}\|^2 \tag{13}$$

subject to

$$y_i \left[\langle \Phi_m(x_i), \mathbf{a} \rangle + b\right] \geq 1 \tag{14}$$

for all $i \in \{1, \ldots, m\}$, where $\mathbf{a} = (a_1, \ldots, a_m)^\top \in \mathbb{R}^m$. In view of (12), however, the constraints (14) are equivalent to $y_i \left[\sum_{j=1}^m a_j K_{ij} + b\right] \geq 1$, i.e. to (7), while the regularizer $\|\mathbf{a}\|^2$ equals (11).

Therefore, using the regularizer (11) and the original kernel essentially[5] corresponds to using a standard SVM with the empirical kernel map. This SVM operates in an m-dimensional feature space with the standard SVM regularizer, i.e., the squared weight of the weight vector in the feature space. We can thus train a classifier using the regularizer (11) simply by using an SVM with the kernel

$$k_m(x, x') := \langle \Phi_m(x), \Phi_m(x') \rangle, \tag{15}$$

and thus, by definition of Φ_m, using the Gram matrix

$$K_m = K K^\top, \tag{16}$$

where K denotes the Gram matrix of the original kernel. The last equation shows that when employing the empirical kernel map, it is not necessary to use a positive definite kernel. The reason is that no matter what K is, the Gram matrix $KK^\top$ is always positive definite,[6] which is sufficient for an SVM.

The remainder of the paper is structured as follows. In Section 2, we introduce the problem of large diagonals, followed by our proposed method to handle it (Section 3). Section 4 presents experiments, and Section 5 summarizes our conclusions.

[5] disregarding the positivity constraints (8)

[6] Here, as in (3), we allow for a nonzero null space in our usage of the concept of positive definiteness.

2 Orthogonal Patterns in the Feature Space

An important feature of kernel methods is that the input domain $\mathcal{X}$ does not have to be a vector space. The inputs might just as well be discrete objects such as strings. Moreover, the map Φ might compute rather complex features of the inputs. Examples thereof are polynomial kernels (Boser et al., 1992), where Φ computes all products (of a given order) of entries of the inputs (in this case, the inputs are vectors), and string kernels (Watkins, 2000; Haussler, 1999; Lodhi et al., 2002), which, for instance, can compute the number of common substrings (not necessarily contiguous) of a certain length $n \in \mathbb{N}$ of two strings x, x' in $O(n|x||x'|)$ time. Here, we assume that x and x' are two finite strings over a finite alphabet Σ. For the string kernel of order n, a basis for the feature space consists of the set of all strings of length n, Σ^n. In this case, Φ maps a string x into a vector whose entries indicate whether the respective string of length n occurs as a substring in x. By construction, these will be rather sparse vectors — a large number of *possible* substrings do not occur in a given string. Therefore, the dot product of two *different* vectors will take a value which is much smaller than the dot product of a vector with itself. This can also be understood as follows: any string shares *all* substrings with itself, but relatively few substrings with another string. Therefore, it will typically be the case that we are faced with *large diagonals*. By this we mean that, given some training inputs $x_1, \ldots, x_m$, we have[7]

$$k(x_i, x_i) >> |k(x_i, x_j)| \text{ for } x_i \neq x_j,\ i, j \in \{1, \ldots, m\}. \tag{17}$$

In this case, the associated Gram matrix will have large diagonal elements.[8]

Let us next consider an innocuous application which is rather popular with SVMs: handwritten digit recognition. We suppose that the data are handwritten characters represented by images in $[0, 1]^N$ (here, $N \in \mathbb{N}$ is the number of pixels), and that only a small fraction of the images is ink (i.e. few entries take the value 1). In that case, we typically have $\langle x, x\rangle > \langle x, x'\rangle$ for $x \neq x'$, and thus the polynomial kernel (which is what most commonly is used for SVM handwritten digit recognition)

$$k(x, x') = \langle x, x'\rangle^d \tag{18}$$

satisfies $k(x, x) >> |k(x, x')|$ already for moderately large d — it has large diagonals.

Note that as in the case of the string kernel, one can also understand this phenomenon in terms of the sparsity of the vectors in the feature space. It is

[7] The diagonal terms $k(x_i, x_i)$ are necessarily nonnegative for positive definite kernels, hence no modulus on the left hand side.

[8] In the machine learning literature, the problem is sometimes referred to as *diagonal dominance*. However, the latter term is used in linear algebra for matrices where the absolute value of each diagonal element is greater than the sum of the absolute values of the other elements in its row (or column). Real diagonally dominant matrices with positive diagonal elements are positive definite.

known that the polynomial kernel of order d effectively maps the data into a feature space whose dimensions are spanned by all products of d pixels. Clearly, if some of the pixels take the value zero to begin with, then an even larger fraction of all possible products of d pixels (assuming $d > 1$) will be zero. Therefore, the sparsity of the vectors will increase with d.

In practice, it has been observed that SVMs do not work well in this situation. Empirically, they work much better if the images are scaled such that the individual pixel values are in $[-1, 1]$, i.e., that the background value is -1. In this case, the data vectors are less sparse and thus further from being orthogonal.

Indeed, large diagonals correspond to approximate orthogonality of any two different patterns mapped into the feature space. To see this, assume that $x \neq x'$ and note that due to $k(x, x) >> |k(x, x')|$,

$$\begin{aligned}\cos(\angle(\Phi(x), \Phi(x'))) &= \frac{\langle \Phi(x), \Phi(x') \rangle}{\sqrt{\langle \Phi(x), \Phi(x) \rangle \, \langle \Phi(x'), \Phi(x') \rangle}} \\ &= \frac{k(x, x')}{\sqrt{k(x, x) k(x', x')}} \approx 0\end{aligned}$$

In some cases, an SVM trained using a kernel with large diagonals will *memorize* the data. Let us consider a simple toy example, using X as data matrix and Y as label vector, respectively:

$$X = \begin{pmatrix} 1\,0\,0\,9\,0\,0\,0\,0\,0\,0 \\ 1\,0\,0\,0\,0\,8\,0\,0\,0\,0 \\ 1\,0\,0\,0\,0\,0\,0\,9\,0\,0 \\ 0\,0\,9\,0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,8\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0\,0\,0\,9 \end{pmatrix}, \quad Y = \begin{pmatrix} +1 \\ +1 \\ +1 \\ -1 \\ -1 \\ -1 \end{pmatrix}$$

The Gram matrix for these data (using the linear kernel $k(x, x') = \langle x, x' \rangle$) is

$$K = \begin{pmatrix} 82 & 1 & 1 & 0 & 0 & 0 \\ 1 & 65 & 1 & 0 & 0 & 0 \\ 1 & 1 & 82 & 0 & 0 & 0 \\ 0 & 0 & 0 & 81 & 0 & 0 \\ 0 & 0 & 0 & 0 & 64 & 0 \\ 0 & 0 & 0 & 0 & 0 & 81 \end{pmatrix}.$$

A standard SVM finds the solution $f(x) = \mathrm{sgn}(\langle w, x \rangle + b)$ with

$$w = (0.04, 0, -0.11, 0.11, 0, 0.12, -0.12, 0.11, 0, -0.11)^\top, \quad b = -0.02.$$

It can be seen from the coefficients of the weight vector w that this solution has but memorized the data: all the entries which are larger than 0.1 in absolute value correspond to dimensions which are nonzero only for *one* of the training points. We thus end up with a look-up table. A *good* solution for a linear classifier, on the other hand, would be to just choose the first feature, e.g., $f(x) = \mathrm{sgn}(\langle w, x \rangle + b)$, with $w = (2, 0, 0, 0, 0, 0, 0, 0, 0, 0)^\top$, $b = -1$.

3 Methods to Reduce Large Diagonals

The basic idea that we are proposing is very simple indeed. We would like to use a nonlinear transformation to reduce the size of the diagonal elements, or, more generally, to reduce the dynamic range of the Gram matrix entries. The only difficulty is that if we simply do this, we have no guarantee that we end up with a Gram matrix that is still positive definite. To ensure that it is, we can use methods of functional calculus for matrices. In the experiments we will mainly use a simple special case of the below. Nevertheless, let us introduce the general case, since we think it provides a useful perspective on kernel methods, and on the transformations that can be done on Gram matrices.

Let K be a symmetric $m \times m$ matrix with eigenvalues in $[\lambda_{\min}, \lambda_{\max}]$, and f a continuous function on $[\lambda_{\min}, \lambda_{\max}]$. Functional calculus provides a unique symmetric matrix, denoted by $f(K)$, with eigenvalues in $[f(\lambda_{\min}), f(\lambda_{\max})]$. It can be computed via a Taylor series expansion in K, or using the eigenvalue decomposition of K: If $K = S^\top DS$ (with D diagonal and S unitary), then $f(K) = S^\top f(D)S$, where $f(D)$ is the diagonal matrix with $f(D)_{ii} = f(D_{ii})$.

The convenient property of this procedure is that we can treat functions of symmetric matrices just like functions on $\mathbb{R}$; in particular, we have, for $\alpha \in \mathbb{R}$, and real continuous functions f, g defined on $[\lambda_{\min}, \lambda_{\max}]$,[9]

$$\begin{aligned}(\alpha f + g)(K) &= \alpha f(K) + g(K)\\ (fg)(K) &= f(K)g(K) = g(K)f(K)\\ \|f\|_{\infty,\sigma(K)} &= \|f(K)\|\\ \sigma(f(K)) &= f(\sigma(K)).\end{aligned}$$

In technical terms, the C^*-algebra generated by K is isomorphic to the set of continuous functions on $\sigma(K)$.

For our problems, functional calculus can be applied in the following way. We start off with a positive definite matrix K with large diagonals. We then reduce its dynamic range by elementwise application of a nonlinear function, such as $\varphi(x) = \log(x+1)$ or $\varphi(x) = \mathrm{sgn}(x) \cdot |x|^p$ with $0 < p < 1$. This will lead to a matrix which may no longer be positive definite. However, it is still symmetric, and hence we can apply functional calculus. As a consequence of $\sigma(f(K)) = f(\sigma(K))$, we just need to apply a function f which maps to $\mathbb{R}_0^+$. This will ensure that all eigenvalues of $f(K)$ are nonnegative, hence $f(K)$ will be positive definite. One can use these observations to design the following scheme.

For positive definite K,

1. compute the positive definite matrix $A := \sqrt{K}$
2. reduce the dynamic range of the entries of A by applying an elementwise transformation φ, leading to a symmetric matrix A_φ
3. compute the positive definite matrix $K' := (A_\varphi)^2$ and use it in subsequent processing. The entries of K' will be the "effective kernel," which in this case is no longer given in analytic form.

[9] Below, $\sigma(K)$ denotes the spectrum of K.

Note that in this procedure, if φ is the identity, then we have $K = K'$.

Experimentally, this scheme works rather well. However, it has one downside: since we no longer have the kernel function in analytic form, our only means of evaluating it is to include all test inputs (not the test labels, though) into the matrix K. In other words, K should be the Gram matrix computed from the observations $x_1, \ldots, x_{m+n}$ where $x_{m+1}, \ldots, x_{m+n}$ denote the test inputs. We thus need to know the test inputs already during training. This setting is sometimes referred to as *transduction* (Vapnik, 1998).

If we skip the step of taking the square root of K, we can alleviate this problem. In that case, the only application of functional calculus left is a rather trivial one, that of computing the square of K. The $m \times m$ submatrix of K^2 which in this case would have to be used for training then equals the Gram matrix when using the empirical kernel map

$$\Phi_{m+n}(x) = (k(x, x_1), \ldots, k(x, x_{m+n}))^\top. \tag{19}$$

For the purposes of computing dot products, however, this can approximately be replaced by the empirical kernel map in terms of the training examples only, i.e., by (12). The justification for this is that for large $r \in \mathbb{N}$, $\frac{1}{r} \langle \Phi_r(x), \Phi_r(x') \rangle \approx \int_{\mathcal{X}} k(x, x'') k(x', x'')\, dP(x'')$, where P is assumed to be the distribution of the inputs. Therefore, we have $\frac{1}{m} \langle \Phi_m(x), \Phi_m(x') \rangle \approx \frac{1}{m+n} \langle \Phi_{m+n}(x), \Phi_{m+n}(x') \rangle$. Altogether, the procedure then boils down to simply training an SVM using the empirical kernel map in terms of the training examples and the transformed kernel function $\varphi(k(x, x'))$. This is what we will use in the experiments below.[10]

4 Experiments

4.1 Artificial Data

We first constructed a set of artificial experiments which produce kernels exhibiting large diagonals. The experiments are as follows: a string classification problem, a microarray cancer detection problem supplemented with extra noisy features and a toy problem whose labels depend upon hidden variables; the visible variables are nonlinear combinations of those hidden variables.

String Classification We considered the following classification problem. Two classes of strings are generated with equal probability by two different Markov models. Both classes of strings consist of letters from the same alphabet of $a = 20$ letters, and strings from both classes are always of length $n = 20$. Strings from the negative class are generated by a model where transitions from any letter to any other letter are equally likely. Strings from the positive class are generated by a model where transitions from one letter to itself (so the next letter is the same as the last) have probability 0.43, and all other transitions have probability 0.03. For both classes the starting letter of any string is equally likely to be any

[10] For further experimental details, cf. Weston and Schölkopf (2001).

letter of the alphabet. The task then is to predict which class a given string belongs to. To map these strings into a feature space, we used the string kernel described above, computing a dot product product in a feature space consisting of all subsequences of length l. In the present application, the subsequences are weighted by an exponentially decaying factor λ of their full length in the text, hence emphasizing those occurrences which are close to contiguous. A method of computing this kernel efficiently using a dynamic programming technique is described by Lodhi et al. (2002). For our problem we chose the parameters $l = 3$ and $\lambda = \frac{1}{4}$.

We generated 50 such strings and used the string subsequence kernel with $\lambda = 0.25$.[11] We split the data into 25 for training and 25 for testing in 20 separate trials. We measured the success of a method by calculating the mean classification loss on the test sets. Figure 1 shows four strings from the dataset and the computed kernel matrix for these strings[12]. Note that the diagonal entries are much larger than the off-diagonals because a long string has a large number of subsequences that are shared with no other strings in the dataset apart from itself. However, information relevant to the classification of the strings is contained in the matrix. This can be seen by computing the mean kernel value between two examples of the positive class which is equal to 0.0003 ± 0.0011, whereas the mean kernel value between two examples of opposite classes is 0.00002 ± 0.00007. Although the numbers are very small, this captures that the positive class have more in common with each other than with random strings (they are more likely to have repeated letters).

string	class
qqbqqnshrtktfhhaahhh	+ve
abajahnaajjjjiiiittt	+ve
sdolncqniflmmpcrioog	-ve
reaqhcoigealgqjdsdgs	-ve

$$K = \begin{pmatrix} 0.6183 & 0.0133 & 0.0000 & 0.0000 \\ 0.0133 & 1.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.4692 & 0.0002 \\ 0.0000 & 0.0000 & 0.0002 & 0.4292 \end{pmatrix}$$

Fig. 1. Four strings and their kernel matrix using the string subsequence kernel with $\lambda = 0.25$. Note that the diagonal entries are much larger than the off-diagonals because a long string has a large number of subsequences that are shared with no other strings in the dataset apart from itself.

If the original kernel is denoted as a dot product $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$, then we employ the kernel $k(x, y) = \langle \Phi(x), \Phi(y) \rangle^p$ where $0 < p < 1$ to solve the diagonal dominance problem. We will refer to this kernel as a *subpolynomial* one. As this kernel may no longer be positive definite we use the method described in

[11] We note that introducing nonlinearities using an RBF kernel with respect to the distances generated by the subsequence kernel can improve results on this problem, but we limit our experiments to ones performed in the linear space of features generated by the subsequence kernel.

[12] Note, the matrix was rescaled by dividing by the largest entry.

Table 1. Results of using the string subsequence kernel on a string classification problem (top row). The remaining rows show the results of using the subpolynomial kernel to deal with the large diagonal.

kernel method		classification loss
original k, $k(x,y) = \langle \Phi(x), \Phi(y) \rangle$		0.36 ± 0.13
$k_{emp}(x,y) = \langle \Phi(x), \Phi(y) \rangle^p$	p=1	0.30 ± 0.08
	p=0.9	0.25 ± 0.09
	p=0.8	0.20 ± 0.10
	p=0.7	0.15 ± 0.09
	p=0.6	$\mathbf{0.13 \pm 0.07}$
	p=0.5	0.14 ± 0.06
	p=0.4	0.15 ± 0.07
	p=0.3	0.15 ± 0.06
	p=0.2	0.17 ± 0.07
	p=0.1	0.21 ± 0.09

Section 1, employing the empirical kernel map to embed our distance measure into a feature space. Results of using our method to solve the problem of large diagonals is given in Table 1. The method provides, with the optimum choice of the free parameter, a reduction from a loss of 0.36 ± 0.13 with the original kernel to 0.13 ± 0.07 with p=0.6. Although we do not provide methods for choosing this free parameter, it is straight-forward to apply conventional techniques of model selection (such as cross validation) to achieve this goal.

We also performed some further experiments which we will briefly discuss. To check that the result is a feature of kernel algorithms, and not something peculiar to SVMs, we also applied the same kernels to another algorithm, kernel 1-nearest neighbor. Using the original kernel matrix yields a loss of 0.43 ± 0.06 whereas the subpolynomial method again improves the results, using $p = 0.6$ yields 0.22 ± 0.08 and $p = 0.3$ (the optimum choice) yields 0.17 ± 0.07. Finally, we tried some alternative proposals for reducing the large diagonal effect. We tried using Kernel PCA to extract features as a pre-processing to training an SVM. The intuition behind using this is that features contributing to the large diagonal effect may have low variance and would thus be removed by KPCA. KPCA did improve performance a little, but did not provide results as good as the subpolynomial method. The best result was found by extracting 15 features (from the kernel matrix of 50 examples) yielding a loss of 0.23 ± 0.07.

Microarray Data With Added Noise We next considered the microarray classification problem of Alon et al. (1999) (see also Guyon et al. (2001) for a treatment of this problem with SVMs). In this problem one must distinguish between cancerous and normal tissue in a colon cancer problem given the expression of genes measured by microarray technology. In this problem one does not encounter large diagonals, however we augmented the original dataset with extra noisy features to simulate such a problem. The original data has 62 ex-

amples (22 positive, 40 negative) and 2000 features (gene expression levels of the tissues samples). We added a further 10,000 features to the dataset, such that for each example a randomly chosen 100 of these features are chosen to be nonzero (taking a random value between 0 and 1) and the rest are equal to zero. This creates a kernel matrix with large diagonals. In Figure 2 we show the first 4×4 entries of the kernel matrix of a linear kernel before and after adding the noisy features.

The problem is again an artificial one demonstrating the problem of large diagonals, however this time the feature space is rather more explicit rather than the implicit one induced by string kernels. In this problem we can clearly see the large diagonal problem is really a special kind of feature selection problem. As such, feature selection algorithms should be able to help improve generalize ability, unfortunately most feature selection algorithms work on explicit features rather than implicit ones induced by kernels.

Performance of methods was measured using 10 fold cross validation, which was repeated 10 times. Due to the unbalanced nature of the number of positive and negative examples in this data set we measured the error rates using a balanced loss function with the property that chance level is a loss of 0.5, regardless of the ratio of positive to negative examples. On this problem (with the added noise) an SVM using the original kernel does not perform better than chance. The results of using the original kernel and the subpolynomial method are given in Table 2. The subpolynomial kernel leads to a large improvement over using the original kernel. Its performance is close to that of an SVM on the original data without the added noise, which in this case is 0.18 ± 0.15.

Hidden Variable Problem We then constructed an artificial problem where the labels can be predicted by a linear rule based upon some hidden variables. However, the visible variables are a nonlinear combination of the hidden variables combined with noise. The purpose is to show that the subpolynomial kernel is not only useful in the case of matrices with large diagonals: it can also improve results in the case where a *linear* rule already overfits. The data are generated as follows. There are 10 hidden variables: each class $y \in \{\pm 1\}$ is generated by a 10 dimensional normal distribution $N(\mu, \sigma)$ with variance $\sigma^2 = 1$, and mean $\mu = y(0.5, 0.5, \ldots, 0.5)$. We then add 10 more (noisy) features for each example, each generated with $N(0, 1)$. Let us denote the 20-dimensional vector obtained

$$K = \begin{pmatrix} 1.00 & 0.41 & 0.33 & 0.42 \\ 0.41 & 1.00 & 0.17 & 0.39 \\ 0.33 & 0.17 & 1.00 & 0.61 \\ 0.42 & 0.39 & 0.61 & 1.00 \end{pmatrix}, \qquad K' = \begin{pmatrix} 39.20 & 0.41 & 0.33 & 0.73 \\ 0.41 & 37.43 & 0.26 & 0.88 \\ 0.33 & 0.26 & 31.94 & 0.61 \\ 0.73 & 0.88 & 0.61 & 35.32 \end{pmatrix}$$

Fig. 2. The first 4×4 entries of the kernel matrix of a linear kernel on the colon cancer problem before (K) and after (K') adding 10,000 sparse, noisy features. The added features are designed to create a kernel matrix with a large diagonal.

Table 2. Results of using a linear kernel on a colon cancer classification problem with added noise (top row). The remaining rows show the results of using the subpolynomial kernel to deal with the large diagonal.

kernel method	balanced loss
original k, $k(x,y) = \langle x, y \rangle$	0.49 ± 0.05
$k_{emp}(x,y) = \operatorname{sgn}\langle x, y \rangle \cdot \mid \langle x, y \rangle \mid^p$ p=0.95	0.35 ± 0.17
p=0.9	0.30 ± 0.17
p=0.8	0.25 ± 0.18
p=0.7	$\mathbf{0.22 \pm 0.17}$
p=0.6	0.23 ± 0.17
p=0.5	0.25 ± 0.19
p=0.4	0.28 ± 0.19
p=0.3	0.29 ± 0.18
p=0.2	0.30 ± 0.19
p=0.1	0.31 ± 0.18

this wasy for example i as h_i. The visible variables x_i are then constructed by taking all monomials of degree 1 to 4 of h_i. It is known that dot products between such vectors can be computed using polynomial kernels (Boser et al., 1992), thus the dot product between two visible variables is

$$k(x_i, x_j) = (\langle h_i, h_j \rangle + 1)^4 .$$

We compared the subpolynomial method to a linear kernel using balanced 10-fold cross validation, repeated 10 times. The results are shown in Table 3. Again, the subpolynomial kernel gives improved results.

One interpretation of these results is that if we know that the visible variables are polynomials of some hidden variables, then it makes sense to use a subpolynomial transformation to obtain a Gram matrix closer to the one we could compute if we were given the hidden variables. In effect, the subpolynomial kernel can (approximately) extract the hidden variables.

4.2 Real Data

Thrombin Binding Problem In the thrombin dataset the problem is to predict whether a given drug binds to a target site on thrombin, a key receptor in blood clotting. This dataset was used in the KDD (Knowledge Discovery and Data Mining) Cup 2001 competition and was provided by DuPont Pharmaceuticals.

In the training set there are 1909 examples representing different possible molecules (drugs), 42 of which bind. Hence the data is rather unbalanced in this respect. Each example has a fixed length vector of 139,351 binary features (variables) in $\{0, 1\}$ which describe three-dimensional properties of the molecule. An important characteristic of the data is that very few of the feature entries are nonzero (0.68% of the 1909×139351 training matrix, see (Weston et al., 2002) for

Table 3. Results of using a linear kernel on the hidden variable problem (top row). The remaining rows show the results of using the subpolynomial kernel to deal with the large diagonal.

kernel method		classification loss
original k, $k(x,y) = \langle x, y \rangle$		0.26 ± 0.12
$k_{emp}(x,y) = \mathrm{sgn}\langle x, y \rangle \cdot \mid \langle x, y \rangle \mid^p$	p=1	0.25 ± 0.12
	p=0.9	0.23 ± 0.13
	p=0.8	0.19 ± 0.12
	p=0.7	0.18 ± 0.12
	p=0.6	$\mathbf{0.16 \pm 0.11}$
	p=0.5	$\mathbf{0.16 \pm 0.11}$
	p=0.4	$\mathbf{0.16 \pm 0.11}$
	p=0.3	0.18 ± 0.11
	p=0.2	0.20 ± 0.12
	p=0.1	0.19 ± 0.13

further statistical analysis of the dataset). Thus, many of the features somewhat resemble the noisy features that we added on to the colon cancer dataset to create a large diagonal in Section 4.1. Indeed, constructing a kernel matrix of the training data using a linear kernel yields a matrix with a mean diagonal element of 1377.9 ± 2825 and a mean off-diagonal element of 78.7 ± 209. We compared the subpolynomial method to the original kernel using 8-fold balanced cross validation (ensuring an equal number of positive examples were in each fold). The results are given in Table 4. Once again the subpolynomial method provides improved generalization. It should be noted that feature selection and transduction methods have also been shown to improve results, above that of a linear kernel on this problem (Weston et al., 2002).

Table 4. Results of using a linear kernel on the thrombin binding problem (top row). The remaining rows show the results of using the subpolynomial kernel to deal with the large diagonal.

kernel method		balanced loss
original k, $k(x,y) = \langle x, y \rangle$		0.30 ± 0.12
$k_{emp}(x,y) = \langle x, y \rangle^p$	p=0.9	0.24 ± 0.10
	p=0.8	0.24 ± 0.10
	p=0.7	0.18 ± 0.09
	p=0.6	0.18 ± 0.09
	p=0.5	$\mathbf{0.15 \pm 0.09}$
	p=0.4	0.17 ± 0.10
	p=0.3	0.17 ± 0.10
	p=0.2	0.18 ± 0.10
	p=0.1	0.22 ± 0.15

Table 5. Results of using a linear kernel on the Lymphoma classification problem (top row). The remaining rows show the results of using the subpolynomial kernel to deal with the large diagonal.

kernel method		balanced loss
original k, $k(x,y) = \langle x, y \rangle$		0.043 ± 0.08
$k_{emp}(x,y) = \operatorname{sgn}\langle x, y \rangle \cdot \lvert \langle x, y \rangle^p$	p=1	0.037 ± 0.07
	p=0.9	0.021 ± 0.05
	p=0.8	0.016 ± 0.05
	p=0.7	$\mathbf{0.015 \pm 0.05}$
	p=0.6	0.022 ± 0.06
	p=0.5	0.022 ± 0.06
	p=0.4	0.042 ± 0.07
	p=0.3	0.046 ± 0.08
	p=0.2	0.083 ± 0.09
	p=0.1	0.106 ± 0.09

Lymphoma Classification We next looked at the problem of identifying large B-Cell Lymphoma by gene expression profiling (Alizadeh et al., 2000). In this problem the gene expression of 96 samples is measured with microarrays to give 4026 features. Sixty-one of the samples are in classes "DLCL", "FL" or "CLL" (malignant) and 35 are labelled "otherwise" (usually normal). Although the data does not induce a kernel matrix with a very large diagonal it is possible that the large number of features induce overfitting even in a linear kernel. To examine if our method would still help in this situation we applied the same techniques as before, this time using balanced 10-fold cross validation, repeated 10 times, and measuring error rates using the balanced loss. The results are given in Table 5. The improvement given by the subpolynomial kernel suggests that overfitting in linear kernels when the number of features is large may be overcome by applying special feature maps. It should be noted that (explicit) feature selection methods have also been shown to improve results on this problem, see e.g Weston et al. (2001).

Protein Family Classification We then focussed on the problem of classifying protein domains into superfamilies in the Structural Classification of Proteins (SCOP) database version 1.53 (Murzin et al., 1995). We followed the same problem setting as Liao and Noble (2002): sequences were selected using the Astral database (astral.stanford.edu cite), removing similar sequences using an E-value threshold of 10^{-25}. This procedure resulted in 4352 distinct sequences, grouped into families and superfamilies. For each family, the protein domains within the family are considered positive test examples, and the protein domains outside the family but within the same superfamily are taken as positive training examples. The data set yields 54 families containing at least 10 family members (positive training examples). Negative examples are taken from outside of the positive sequence's fold, and are randomly split into train and test sets in

the same ratio as the positive examples. Details about the various families are listed in (Liao and Noble, 2002), and the complete data set is available at www.cs.columbia.edu/compbio/svm-pairwise. The experiments are characterized by small positive (training and test) sets and large negative sets. Note that this experimental setup is similar to that used by Jaakkola et al. (2000), except the positive training sets do not include additional protein sequences extracted from a large, unlabeled database, which amounts to a kind of "transduction" (Vapnik, 1998) algorithm.[13]

An SVM requires fixed length vectors. Proteins, of course, are variable-length sequences of amino acids and hence cannot be directly used in an SVM. To solve this task we used a sequence kernel, called the spectrum kernel, which maps strings into a space of features which correspond to every possible k-mer (sequence of k letters) with at most m mismatches, weighted by prior probabilities (Leslie et al., 2002). In this experiment we chose $k = 3$ and $m = 0$. This kernel is then normalized so that each vector has length 1 in the feature space; i.e.,

$$k(x, x') = \frac{\langle x, x' \rangle}{\sqrt{\langle x, x \rangle \langle x', x' \rangle}}. \tag{20}$$

An asymmetric soft margin is implemented by adding to the diagonal of the kernel matrix a value $0.02 \cdot \rho$, where ρ is the fraction of training set sequences that have the same label as the current sequence (see Cortes and Vapnik (1995); Brown et al. (2000) for details). For comparison, the same SVM parameters are used to train an SVM using the Fisher kernel (Jaakkola and Haussler (1999); Jaakkola et al. (2000), see also Tsuda et al. (2002)), another possible kernel choice. The Fisher kernel is currently considered one of the most powerful homology detection methods. This method combines a generative, profile hidden Markov model (HMM) and uses it to generate a kernel for training an SVM. A protein's vector representation induced by the kernel is its gradient with respect to the profile hidden Markov model, the parameters of which are found by expectation-maximization.

For each method, the output of the SVM is a discriminant score that is used to rank the members of the test set. Each of the above methods produces as output a ranking of the test set sequences. To measure the quality of this ranking, we use two different scores: receiver operating characteristic (ROC) scores and the median rate of false positives (RFP). The ROC score is the normalized area under a curve that plots true positives as a function of false positives for varying classification thresholds. A perfect classifier that puts all the positives at the top of the ranked list will receive an ROC score of 1, and for these data, a random classifier will receive an ROC score very close to 0. The median RFP score is the fraction of negative test sequences that score as high or better

[13] We believe that it is this transduction step which may be responsible for much of the success of using the methods described by Jaakkola et al. (2000)). However, to make a fair comparison of kernel methods we do not include this step which could potentially be included in any of the methods. Studying the importance of transduction remains a subject of further research.

Table 6. Results of using the spectrum kernel with $k = 3, m = 0$ on the SCOP dataset (top row). The remaining rows (apart from the last one) show the results of using the subpolynomial kernel to deal with the large diagonal. The last row, for comparison, shows the performance of an SVM using the Fisher kernel.

kernel method		RFP	ROC
original k, $k(\Phi(x), \Phi(y)) = \langle x, y \rangle$		0.1978	0.7516
$k_{emp}(x, y) = \langle \Phi(x), \Phi(y) \rangle^p$	p=0.5	0.1697	0.7967
	p=0.4	0.1569	0.8072
	p=0.3	0.1474	0.8183
	p=0.2	**0.1357**	**0.8251**
	p=0.1	0.1431	0.8213
	p=0.05	0.1489	0.8156
SVM-FISHER		0.2946	0.6762

than the median-scoring positive sequence. RFP scores were used by Jaakkola *et al.* in evaluating the Fisher-SVM method. The results of using the spectrum kernel, the subpolynomial kernel applied to the spectrum kernel and the fisher kernel are given in Table 6. The mean ROC and RFP scores are superior for the subpolynomial kernel. We also show a family-by-family comparison of the subpolynomial spectrum kernel with the normal spectrum kernel and the Fisher kernel in Figure 3. The coordinates of each point in the plot are the ROC scores for one SCOP family. The subpolynomial kernel uses the parameter $p = 0.2$. Although the subpolynomial method does not improve performance on every single family over the other two methods, there are only a small number of cases where there is a loss in performance.

Note that explicit feature selection cannot readily be used in this problem, unless it is possible to integrate the feature selection method into the construction of the spectrum kernel, as the features are never explicitely represented. Thus we do not know of another method that can provide the improvements described here. Note though that the improvements are not as large as reported in the other experiments (for example, the toy string kernel experiment of Section 4.1). We believe this is because this application does not suffer from the large diagonal problem as much as the other problems. Even without using the subpolynomial method, the spectrum kernel is already superior to the Fisher kernel method. Finally, note that while these results are rather good, they do not represent the record results on this dataset: in (Liao and Noble, 2002), a different kernel (Smith-Waterman pairwise scores)[14] is shown to provide further improvements (mean RFP: 0.09, mean ROC: 0.89). It is also possible to choose other parameters of the spectrum kernel to improve its results. Future work will continue to investigate these kernels.

[14] The Smith-Waterman score technique is closely related to the empirical kernel map, where the (non-positive definite) effective "kernel" is the Smith-Waterman algorithm plus p-value computation.

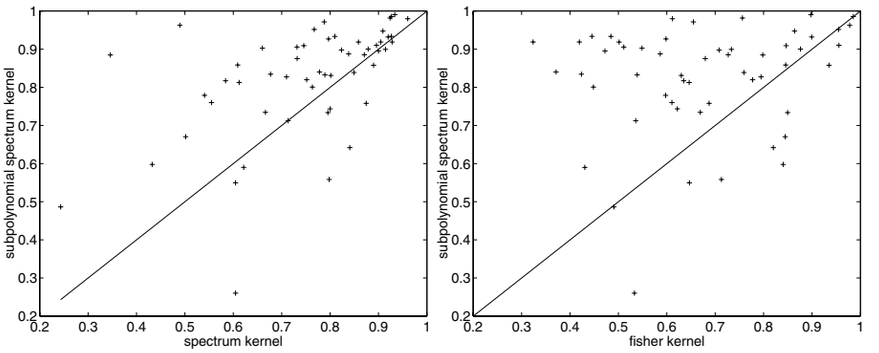

Fig. 3. Family-by-family comparison of the subpolynomial spectrum kernel with: the normal spectrum kernel (left), and the Fisher kernel (right). The coordinates of each point in the plot are the ROC scores for one SCOP family. The spectrum kernel uses $k = 3$ and $m = 0$, and the subpolynomial kernel uses p=0.2. Points above the diagonal indicate problems where the subpolynomial kernel performs better than the other methods.

5 Conclusion

It is a difficult problem to construct useful similarity measures for non-vectorial data types. Not only do the similarity measures have to be positive definite to be useable in an SVM (or, more generally, conditionally positive definite, see e.g. Schölkopf and Smola (2002)), but, as we have explained in the present paper, they should also lead to Gram matrices whose diagonal values are not overly large. It can be difficult to satisfy both needs simultaneously, a prominent example being the much celebrated (but so far not too much used) string kernel. However, the problem is not limited to sophisticated kernels. It is common to all situations where the data are represented as sparse vectors and then processed using an algorithm which is based on dot products.

We have provided a method to deal with this problem. The method's upside is that it turns kernels such as string kernels into kernels that work very well on real-world problems. Its main downside so far is that the precise role and the choice of the function we apply to reduce the dynamic range has yet to be understood.

Acknowledgements We would like to thank Olivier Chapelle and André Elisseeff for very helpful discussions. We moreover thank Chris Watkins for drawing our attention to the problem of large diagonals.

Bibliography

A. A. Alizadeh et al. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000. Data available from http://llmpp.nih.gov/lymphoma.

U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon cancer tissues probed by oligonucleotide arrays. *Cell Biology*, 96:6745–6750, 1999.

C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer-Verlag, New York, 1984.

B. E. Boser, I. M. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, PA, July 1992. ACM Press.

M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. Sugnet, T. S. Furey, M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data using support vector machines. *Proceedings of the National Academy of Sciences*, 97(1):262–267, 2000.

C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20: 273–297, 1995.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 2001.

D. Haussler. Convolutional kernels on discrete structures. Technical Report UCSC-CRL-99-10, Computer Science Department, University of California at Santa Cruz, 1999.

T. S. Jaakkola, M. Diekhans, and D. Haussler. A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 7: 95–114, 2000.

T. S. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11*, Cambridge, MA, 1999. MIT Press.

C. Leslie, E. Eskin, and W. S. Noble. The spectrum kernel: A string kernel for SVM protein classification. *Proceedings of the Pacific Symposium on Biocomputing*, 2002. To appear.

L. Liao and W. S. Noble. Combining pairwise sequence similarity and support vector machines for remote protein homology detection. *Proceedings of the Sixth International Conference on Computational Molecular Biology*, 2002.

H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2: 419–444, 2002.

A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, pages 247:536–540, 1995.

E. Osuna and F. Girosi. Reducing the run-time complexity in support vector machines. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 271–284, Cambridge, MA, 1999. MIT Press.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

K. Tsuda. Support vector classifier with asymmetric kernel function. In M. Verleysen, editor, *Proceedings ESANN*, pages 183–188, Brussels, 1999. D Facto.

K. Tsuda, M. Kawanabe, G. Rätsch, S. Sonnenburg, and K.R. Müller. A new discriminative kernel from probabilistic models. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2002. To appear.

V. Vapnik. *Estimation of Dependences Based on Empirical Data [in Russian]*. Nauka, Moscow, 1979. (English translation: Springer Verlag, New York, 1982).

V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.

C. Watkins. Dynamic alignment kernels. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 39–50, Cambridge, MA, 2000. MIT Press.

J. Weston, A. Elisseeff, and B. Schölkopf. Use of the ℓ_0-norm with linear models and kernel methods. *Biowulf Technical report*, 2001. http://www.conclu.de/∼jason/.

J. Weston, F. Pérez-Cruz, O. Bousquet, O. Chapelle, A. Elisseeff, and B. Schölkopf. Feature selection and transduction for prediction of molecular bioactivity for drug design, 2002. http://www.conclu.de/~jason/kdd/kdd.html.

J. Weston and B. Schölkopf. Dealing with large diagonals in kernel matrices. In *New Trends in Optimization and Computational algorithms (NTOC 2001), Kyoto, Japan*, 2001.