SIGNAL THEORY FOR SVM KERNEL PARAMETER ESTIMATION

J. D. B. Nelson, R. I. Damper, S. R. Gunn and B. Guo

Information: Signals, Images, Systems Research Group
School of Electronics and Computer Science
University of Southampton, Southampton SO17 1BJ, UK

ABSTRACT

Fourier-based regularisation is considered for the support vector machine classification problem over absolutely integrable loss functions. By invoking the modest assumption that the decision function belongs to a Paley-Wiener space, it is shown that the classification problem can be developed in the context of signal theory. Furthermore, by employing the Paley-Wiener reproducing kernel, namely the sinc function, it is shown that a principled and finite kernel hyper-parameter search space can be discerned, a priori. Subsequent experiments, performed on a commonly available hyper-spectral image data set, reveal that the approach yields results that surpass state-of-the-art benchmarks.

1. INTRODUCTION

Parameter choice is an open problem in support vector machine (SVM) learning. Whether the parameter takes the form of a scaling vector, a scaling number, or the kernel itself, the fact remains that in the context of non-linear support vector machines there are uncountably many solutions. Unfortunately, the only way to elicit the best solution is to build uncountably many kernels. This is, of course, intractable.

However, when framed in the context of reproducing kernel Hilbert spaces, it can be shown that the parameters control the nature and degree of regularisation that is imposed on the solution. A related issue is that the so-called curse of dimensionality often turns out to be unexpectedly ineffectual. Some recent machine learning research has focused on finding cogent explanations for this phenomenon. Belkin and Niyogi [1] argue that a possible reason is that the data lie on a sub-manifold, embedded in the input space. Indeed, data with a large number of variables may lie entirely in a much smaller dimensional manifold. Knowledge pertaining to the structure of the manifold can be used to guide the choice of parameters, and thus the nature and degree of regularisation. Such realisations lead to a more considered approach: that is to ascertain, a priori, properties of the space wherein the data lies. Although there may still exist infinitely many solutions, the range of an empirical search could then at least be focused upon subsets of parameters rather than all possible choices of parameters. In fact, we propose principled assertions that reduce the infinite search space to a finite one. Ultimately, our philosophy is inspired by the discipline of sampling theory where the main goal is to establish equivalence relations between data sequence spaces and kernel function spaces. To this end, we employ perhaps the most simple function space from sampling theory, namely the simply connected and zero-centred Paley-Wiener reproducing kernel Hilbert space, more commonly referred to by engineers as base-band-limited signals. For a given class of data we show how to estimate, a priori, a suitable kernel and parameter subspace.

The remainder of this paper is structured as follows. In Section 2 the data class and corresponding reproducing kernel Hilbert space are constructed. Accordingly, some necessary signal theory concepts are introduced and discussed in Section 3, and exploited in Section 4. Finally, in Section 5, we announce the best results to date on a popular hyper-spectral image data set.

This research was supported by the Data Information Fusion Defence Technology Centre, United Kingdom, under DTC Project 8.2.

2. MODEL CONSTRUCTION

Consider the usual SVM classification problem, with $x_n \in X \subseteq \mathbb{R}^d, y_n \in \{\pm 1\}$, $n \in \mathbb{N}$,

$$\min_{f \in \mathcal{H}} \frac{1}{2} \|\Gamma f\|^2 + C \sum_{n \in \mathbb{N}} |1 - y_n f(x_n)|_+ ,$$

where f, the decision function to be determined, in some Hilbert space $\mathcal{H}(X)$, is regularised by the operator $\Gamma \colon \mathcal{H} \mapsto \mathcal{F}$. The resulting learned decision function, implied by the representor theorem [2], is the solution $f = \sum_{n \in \mathbb{N}} y_n \alpha_n k(x_n, \cdot)$, where k is a Mercer kernel [3]. Herewith, the classifier is defined by $\operatorname{sgn} f$. Our main contention here is that before an effort is made to solicit the classifier it is good practice, in a qualitative sense, to attempt to discern the properties of the underlying decision function. A natural preface, proposed in this work, is that the labelling function maps d-variate data to labels via $y\colon \mathbb{R}^d \supset X \mapsto \{\pm 1\}$, with

$$y(x) := \operatorname{sgn}\big(\varphi(x) + \varepsilon(x)\big), \tag{1}$$

where the noise is modelled by ε, and under the assumption that the information content φ, lies entirely within the space of Paley-Wiener (PW) functions over some multi-dimensional base-band region Ω^*, viz.

$$\varphi \in PW_{\Omega^*} := \bigoplus_{r=1}^{d} \left\{\zeta \in L_2(X) : \operatorname{supp} \zeta^\wedge \subseteq \Omega_r^*\right\}, \tag{2}$$

with $\operatorname{supp} \zeta := \{x \in X : \zeta(x) \neq 0\}$, and where $\cdot^\wedge$ denotes d-variate Fourier transformation. The condition $\varphi \in PW_{\Omega^*}$ restricts the behaviour of the information content to functions of finite bandwidth around the origin. Although this kernel is familiar to signal theorists and engineers, it is a seemingly rare tool in machine learning. It is, perhaps less well known that, by virtue of the following three established results, the sinc kernel also lends itself to the regularised support vector classification setting.

Theorem 2.1 *(Self consistency property, Smola, Schölkopf, and Müller [4]) Let the Mercer kernel $k\colon X \times X \mapsto \mathbb{R}$, and the regularisation operator $\Gamma\colon \mathcal{H} \mapsto \mathcal{F}$, be such that $k(x,\xi) \equiv \big\langle (\Gamma k)(x), (\Gamma k)(\xi) \big\rangle_{\mathcal{F}}$. Then the SVM classification problem can be written*

$$\min_{f \in \mathcal{H}} \frac{1}{2} \|\Gamma f\|^2 + C \sum_{n \in \mathbb{N}} |1 - y_n f(x_n)|_+ .$$

Theorem 2.2 *(Translation invariant kernels, Smola, Schölkopf, and Müller [4]) Consider a kernel, endowed with translation invariance, namely $k(x,\xi) = k(x-\xi)$, with the regularisation operator $\Gamma\colon \mathcal{H} \mapsto \mathcal{F}$, defined by*

$$\langle \Gamma f, \Gamma g \rangle_{\mathcal{F}} = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \frac{f^\wedge(\omega)\overline{g^\wedge(\omega)}}{k^\wedge(\omega)} \, \mathrm{d}\omega.$$

Then $k(x,\xi) \equiv \big\langle (\Gamma k)(x), (\Gamma k)(\xi) \big\rangle_{\mathcal{F}}$, and the self consistency property from Theorem 2.1 is satisfied.

Corollary 2.3 *It follows from Theorem 2.2 that the regularisation term from the SVM problem is*

$$\|\Gamma f\|_{\mathcal{F}}^2 = \frac{1}{(2\pi)^{d/2}} \prod_{r=1}^{d} \int_{\Omega_r^*} \frac{|f^\wedge(\omega)|^2}{k_r^\wedge(\omega^r)} \, \mathrm{d}\omega^r,$$

with $\omega := (\omega^r)_{r=1}^d$, and that $(k^\wedge(\omega))^{-1} = \left(\prod_{r=1}^d k_r^\wedge(\omega^r)\right)^{-1}$ regularises the decision function f by acting as a filter, in the signal analysis sense, on $|f^\wedge|^2$.

The unique kernel associated with the reproducing kernel Hilbert space PW_{Ω^*} is the sinc kernel $\prod_{r=1}^d \operatorname{sinc}_{\omega_*^r}(x^r - \xi^r)$. Given the model (1), where the information content is embedded in the Paley-Wiener space (2), it is only sensible to constrain the decision function to the same Paley-Wiener space. From Corollary 2.3, it follows that in the Fourier domain the multiplicative filter that acts upon $|f^\wedge|^2$ is

$$\frac{1}{k^\wedge(\omega)} = \frac{1}{\chi_{\Omega^*}(\omega)} = \prod_{r=1}^{d} \frac{1}{\chi_{\Omega_r^*}(\omega^r)} \ ,$$

with the d-dimensional hypercuboid

$$\chi_\Omega(\omega) := \begin{cases} 1 & \text{if } \omega \in \Omega \\ 0 & \text{otherwise} \end{cases} .$$

In this case, since $k^\wedge \geq 0$ holds over $\mathbb{R}^d$, Bochner's theorem ensures that the sinc kernel is a Mercer kernel. The multiplicative filter regularises the decision function by penalising the frequency content of f on $\mathbb{R} \backslash \Omega^*$. The sinc kernel also keeps the content over Ω^* unaltered. These penalisation and preservation properties are, by definition, unique to the sinc kernel. Since Paley-Wiener spaces are closed under addition, the representor result ensures that the decision function is restricted to PW_{Ω^*}.

Remark 2.4 *We now see that, in the context of our work, the non-regularised, higher dimensional input space discussed by Belkin and Niyogi [1] is $PW_{\mathbb{R}^d}$, and the sub-manifold is $PW_{\Omega^*} \subseteq PW_{\mathbb{R}^d}$. That is, in the frequency domain, the sub-manifold invoked by our work can be described as a hypercuboid centred on the origin, and the regularising operator is precisely the mapping $\Gamma\colon PW_{\mathbb{R}^d} \mapsto PW_{\Omega^*}$.*

We are now left with the problem of finding an optimal hyper-parameter set $\{\omega_*^r\}$, in the sense of the SVM problem. Before this is attempted, we propose a novel approach to elicit spectral properties of the labelling function that employs some recently constructed tools from signal theory.

3. FROM SIGNAL THEORY TO SVM CLASSIFICATION

Intuitively, the labelling function y can be understood as a piecewise constant function that maps d-many real variables to positive, or negative, unity. It can, therefore, be treated as a square wave function over d-variate space. To this end, we propose the use of sequency analysis as a means to elicit some properties of y and, consequently, the information content φ. Such properties will suggest how the decision function should be regularised. Before the analysis, it is instructive to introduce a family of functions that has the labelling function as a member.

Let $\mathrm{cal}_\omega(t) := \operatorname{sgn}\cos\omega t$, and $\mathrm{sal}_\omega(t) := \operatorname{sgn}\sin\omega t$, and define the complex square wave family as

$$\psi_\omega := \sqrt{\frac{\pi}{32}}\left(\mathrm{cal}_\omega + i\,\mathrm{sal}_\omega\right).$$

This differs from the definition of the more common Walsh-Hadamard analysis described elsewhere. In particular, the system employed here is defined over a denser, uniform grid rather than over a dyadic grid and, as will be shown below, it forms a biorthogonal basis. As such, it can be used to analyse the spectral properties of functions over a more opaque domain. Consider the Möbius arithmetic function $\mu\colon \mathbb{N} \mapsto \{0, \pm 1\}$, given by

$$\mu(n) := \begin{cases} 1, & \text{if } n = 1 \\ (-1)^m, & \text{if } n \text{ is the product of } m \text{ distinct primes} \\ 0, & \text{otherwise} \end{cases}$$

which is employed here due to the utility afforded by the following result, taken from number theory.

Lemma 3.1 *(Möbius) Let μ denote the Möbius function. Then, for $m \in \mathbb{N}$,* $$\sum_{n|m} \mu(n) = \delta_{m,1},$$
where the $\delta_{\cdot,\cdot}$ denotes the Kronecker delta. The next result, outlined by Nelson [5], enables us to express the labelling function in terms of the complex square wave family.

Proposition 3.2 *(Biorthogonal complex square wave system, Nelson [5]) The biorthogonal dual of $\{\psi_n\}$ is*

$$\psi_n^*(t) := \frac{1}{\sqrt{2\pi}} \sum_{m\in 4\mathbb{Z}+1} m^{-1}\mu(|m|)e^{int/m}.$$

We introduce the sequency transformation, $\cdot^\sim$, namely

$$f^\sim(\omega) = \int_{\mathbb{R}} f(t)\overline{\psi_\omega^*(t)}\,\mathrm{d}t. \quad (3)$$

From Proposition 3.2, it follows that y can be expanded as a superposition of square waves, viz.

$$y = \sum_{n\in\mathbb{Z}} \langle y, \psi_n^* \rangle_{L_2(\mathbb{R})} \psi_n.$$

Hence, the coefficients that express y in terms of the square wave basis are found by performing the sequency transform of y. Recall from (1) that $\varphi \in PW_{\Omega^*}$, and, without loss of generality, $\varepsilon \in PW_{\Omega^+}$. The linearity property of Paley-Wiener spaces gives rise to

$$\varphi + \varepsilon \in PW_{\Omega^* \cup \Omega^+}.$$

We define the sequency function space S_Ω as

$$S_\Omega := \{\zeta \in L_2(X) : \operatorname{supp}\zeta^\sim \subseteq \Omega\},$$

Now since $\varphi \in PW_{\Omega^*} \Rightarrow \operatorname{sgn}\varphi \in S_{\Omega^*}$, and $\varepsilon \in PW_{\Omega^+} \Rightarrow \operatorname{sgn}\varepsilon \in S_{\Omega^+}$, we can express the labelling function y, as a sequency-limited function, $y = \operatorname{sgn}(\varphi + \varepsilon) \in S_{\Omega^* \cup \Omega^+}$, with

$$y = \int_{\Omega^* \cup \Omega^+} y^\sim(\omega)\psi_\omega(\cdot)\,\mathrm{d}\omega, \quad (4)$$

and where $y^\sim$ can be computed via

$$\begin{aligned} y^\sim(\omega^r) &= \frac{1}{\sqrt{2\pi}} \sum_{m\in 4\mathbb{Z}+1} \frac{\mu(|m|)}{m} \int_{\mathbb{R}} y(t)e^{-i\omega x^r/m}\,\mathrm{d}t \\ &= \sum_{m\in 4\mathbb{Z}+1} \frac{\mu(|m|)}{m} y^\wedge\left(\frac{\omega^r}{m}\right), \end{aligned} \quad (5)$$

where one fast Fourier transform is required to determine $y^\wedge(\omega^r)$, for each $r = 1, \ldots, d$. Since the samples x_n^r, over which the Fourier transforms of $y^\wedge(\omega^r)$ are computed, are typically non-uniformly distributed, the direct application of a Fourier transform is inappropriate. Instead, irregular sampling techniques must be considered. Since a comprehensive treatment of irregular sampling issues is beyond the scope of this work, we employ a simple strategy whereby the data are mapped to a uniform grid via nearest neighbour, constant interpolation.

By definition, the information content of $\varphi + \varepsilon$ lies in the frequency base-band $\Omega^* = (-\omega_*\pi, \omega_*\pi)$. Analogously, the informative part of the labelling function $\operatorname{sgn}(\varphi + \varepsilon)$ lies inside some sequency base-band $\Omega^* = (-\omega_*\pi, \omega_*\pi)$.

Example 3.3 *Consider $y = \operatorname{sgn}\varphi$, where $\varphi(t) = \cos\omega_* t$, and $t \in \mathbb{R}$. Clearly, it follows that $\varphi \in PW_{(-\omega_*, \omega_*)}$, and*

$$y^\sim(\omega) = \delta(\omega - \omega_*) + \delta(\omega + \omega_*) \Rightarrow y \in S_{(-\omega_*, \omega_*)}.$$

In this case, ω_* is estimated from $y^\sim$, and $\operatorname{sinc}(\omega_* \cdot)$ is chosen as the kernel.

In practice, the approach taken to determine Ω^*, and hence the value of ω_*, is not straightforward unless we assume that $\Omega^* \cap \Omega^+ = \{\}$. However, in this section we have formulated the SVM classification problem in terms of a signal theory one, namely that of filter design, and in Section 4 we show how this avoids the necessity of unduly repeated implementation of computationally expensive parameter estimators such as cross validation.

4. PARAMETER ESTIMATION

For each choice of the parameter ω_*, there is a corresponding reproducing kernel Hilbert space $\mathcal{H}_*$, say. Commonly, the choice of parameter, or hyper-parameter, is achieved by estimating the performance of the SVM for each parameter value. The value that yields the best performance is then chosen as the optimal parameter.

There exist several different ways to measure SVM performance. To expedite the empirical comparisons, drawn in Section 5, we shall consider perhaps the most straightforward measure, namely the validation error. Here, the data are split into two distinct sets. One set is used to train and the other to validate the SVM.

There also exist several ways to search for the optimal parameter, ω_*. Often misused, the phrase 'exhaustive search' has been adopted to describe an approach whereby the performance measure is computed over a finite number of parameters. In practice, however, the search can never be truly exhaustive. Either the range of parameters is too small, or the discretisation too large, or both. Various gradient descent search methods have also been applied to SVM parameter optimisation. Common drawbacks of gradient methods include finding a suitable smoothing strategy for the performance measure, choosing a good first initial point, and bad convergence.

Unfortunately, the inherent problems of any search-based method are exacerbated in an exponential manner as the number of parameters increase linearly, and when using a one-against-one strategy for example, in a combinatorial manner as the number of classes increase linearly. Only a few authors have attempted automatic estimation of the optimal hyper-parameter set. Lanckriet et al. [6] use semi-definite programming techniques to compute the kernel matrix. Debnath and Takahashi [7] attempt to link the eigenvalues of the features and the optimal Gaussian parameter. However, their work relies almost entirely on empirical evidence and qualitative remarks. Guo et al. use mutual information theory to guide parameter scaling [8].

We propose a principled means to estimate a search space wherein the optimal parameter lies. Rather than blindly searching for a set of parameters by induction alone, we follow an approach that is inspired by the filter design engineering discipline. Although filter design is sometimes glibly described as 'more of an art than a science', it has a successful theoretical and practical history that arguably stretches further back than statistical machine learning. Not only does signal theory suggest parameters a priori, it can also, via spectral analysis, aid the interpretation of the underlying properties of a particular solution.

Our approach is to compute the sequency transform (3), via the series of fast Fourier transforms (5), so as to discern the interval Ω^*, from Equation (4). For a d-variate space $\Omega = \bigoplus_1^d \Omega_r$, we require d-many sequency transforms. When $\Omega_r^* = (-\omega_*^r\pi, \omega_*^r\pi)$ has been established, we use the estimate ω_*^r to construct the sinc kernel under the assumption that $\Omega^* \cap \Omega^+ = \{\}$. In practice, because each datum has finite length, the sequency transform (3) is taken over a finite domain T. From Equation (5) and the convolution theorem this is equivalent to computing

$$(\chi_T y)^{\sim}(\omega) = \frac{T}{2\pi} \sum_{m \in 4\mathbb{Z}+1} \frac{\mu(|m|)}{m} \left(\mathrm{sinc}_T * y^{\wedge}\right)\left(\frac{\omega}{m}\right),$$

where $*$ denotes the convolution operator. Consequently, like the finite Fourier transform, the finite sequency transform is subject to so called sinc ringing effects.

Notwithstanding such artefacts, the sequency components can still be estimated. The shifted Dirac generalised functions found in the idealised and trivial Example 3.3 above are replaced by shifted sinc functions in the finite case. It follows that only the locations of the local maxima of $|y^{\sim}|$ should be considered as candidates for ω_*. Since y is necessarily restricted to a discrete and finite domain, the sequency spectrum is smooth and cannot take the same value at every point. Hence, only finitely many maxima will exist. This simple and intuitive argument serves to reduce an exhaustive but theoretically infinite search to an exhaustive, finite search. For a one-dimensional problem, one merely tests the performance of the SVM by setting the parameter value to each local maximum of the sequency spectrum. Of course, when the number of dimensions or maxima preclude the tractability of an exhaustive search over the entire set, one may be compelled to compromise accuracy and either bound the search space, conduct a sparser search, or both. To this end, we facilitate a disciplined compromise between search sparsity and accuracy by the following construct.

Definition 4.1 *Define the sequency transform of y over the rth variate x^r, by $y^{\sim}(\omega^r)$. The sequence $\{\omega_p^r\}_{p=1}^{P_r}$ is defined as the set that contains the locations of the local maxima of $|y^{\sim}(\omega^r)|$, ordered such that $\omega_p^r \leq \omega_{p+1}^r$, for all $p_r = 1,\ldots,P_r$. Furthermore, define the sets*

$$W_1(\kappa) := \{\omega_1^r\}_{r=1}^d,$$

and

$$W_j(\kappa) := M_j^{\uparrow}(\kappa) \cup W_{j-1}(\kappa) \setminus M_j(\kappa),$$

with

$$M_j(\kappa) := \{\omega_{s_r}^r \in W_{j-1}(\kappa) \colon \omega_{s_r}^r - \min W_{j-1}(\kappa) < \kappa\},$$

and where the set operator $\cdot^{\uparrow}$ is defined as

$$M_j^{\uparrow} \colon M_j = \{\omega_{s_r}^r\} \mapsto \{\omega_{s_r+1}^r\}.$$

Example 4.2 *Consider the set $W_1(0) := \{\omega_1^r\}_{r=1}^3$, with $\omega_1^1 < \omega_1^2 < \omega_2^1 < \omega_1^3$. It follows that $M_2(0) = \{\omega_1^1\}$, $M_2^{\uparrow}(0) = \{\omega_2^1\}$, and $W_2(0) = \{\omega_2^1, \omega_1^2, \omega_1^3\}$. Likewise, we have $W_3(0) = \{\omega_2^1, \omega_2^2, \omega_1^3\}$, and $W_4(0) = \{\omega_3^1, \omega_2^2, \omega_1^3\}$.*

The set $\{W_j(\kappa)\}_j$ is a subset of points that lie in the set of all sequency maxima. It is constructed such that a search over this subspace is not unduly influenced by the sequency spectrum of any one particular dimension relative to the other $d-1$ dimensions. Equivalently, it assumes that the spectral bandwidth of the noise, or information, does not change too much from one dimension to another. Larger κ produce

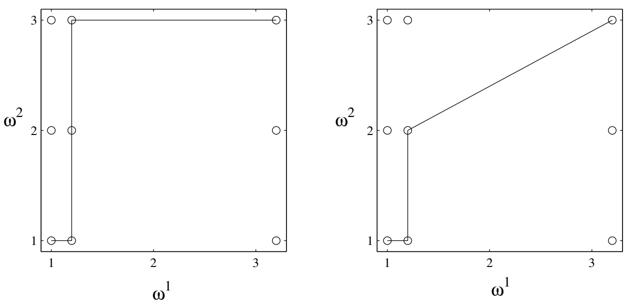

Fig. 1. The circles denote the location of the maxima over a 2-dimensional domain. The lines plot the searches $\{W_j(0)\}_{j=1}^{5}$ on the left and $\{W_j(0.2)\}_{j=1}^{4}$ on the right.

sparser search sets. Figure 1 depicts a simple example for two different values of κ. Herewith lies a useful compromise between accuracy and sparsity. The result is a family of search spaces parameterised by κ which should be chosen in accordance with the computational resources available.

5. APPLICATION TO HYPER-SPECTRAL IMAGERY

The airborne visual and infrared imaging system (AVIRIS) hyper-spectral image data comprises intensity information over 224 coterminous electromagnetic spectral bands, ranging from 0.4 to 2.5 μm. AVIRIS data facilitate myriad applications including resource management, mineral exploitation, and environmental monitoring. The large number of variables, and classes, makes the data set ideal for demonstrating the utility of our sinc kernel approach and search strategy. Furthermore, there exists a free and publicly available AVIRIS data set [9], that has been used by several research groups to benchmark various hyper-spectral image classification techniques. The following experiments make use of this data.

In the hyper-spectral image context, each pixel is described by a single data point, $x_n \in \mathbb{R}^d$. Each element x_n^r, represents the intensity value of pixel n, in the rth spectral band. Each pixel belongs to one of seventeen different classes of ground vegetation. Previous work on the data set has considered four-, sixteen-, and seventeen-class problems. For a fair comparison to be drawn between our results and others, we follow the same sampling and validation technique exercised by previous research on the AVIRIS data. That is, 20% of the original data is randomly chosen as training data, and the remaining 80% is held out as the testing data. The resulting validation measure is simply the percentage of incorrect classifications on the testing data. Figure 2 shows the sequency spectra $y^{\sim}$ taken from the four-class AVIRIS problem.

Table 1 compares results using the proposed sinc methods and the best results found by previous researchers.

Table 1. AVIRIS classification: State-of-the-art

Source	Penalty	Method	Error (%)
Four-class problem			
Section 4, Definition 4.1	∞	**Sinc SVM, sparse search**	**3.9**
Gualtieri and Cromp [10] (5 trials)	1000	SVM poly. kernel, degree 7	4.1
Du [13]	1000	SVM poly. kernel, degree 7	4.5
This work	1000	SVM poly. kernel, degree 7	4.7
This work	∞	Gaussian RBF kernel	4.9
Tadjudin and Landgrebe [11, 12]	1000	Bayesian discrim. analysis	6.5
Du	1000	Gaussian RBF kernel	7.9
Sixteen-class problem			
Section 4, Definition 4.1	∞	**Sinc SVM, sparse search**	**10.9**
Gualtieri and Cromp (1 trial)	1000	SVM poly. kernel, degree 7	12.7
Seventeen-class problem			
Section 4, Definition 4.1	∞	**Sinc SVM, sparse search**	**11.3**
This work	1000	SVM poly. kernel, degree 7	15.1
Tadjudin and Landgrebe	1000	Bayesian discrim. analysis	17.1

Gualtieri and Cromp [10] tested several orders of polynomial SVM kernels over 5 trials and found that the degree-7 kernel performed the best. We can see that the SVM approach holds a significant advantage over the Bayesian method used by Tadjudin [11] and Landgrebe [12].

The sinc-based search strategy implemented here is the sparse hyper-parameter search space $\{W_j(0.05)\}_{j=1}^{5}$ from Definition 4.1. All of the sinc kernel results represent the average, taken over 10 trials. The mean standard error was below 0.2% for the four-class problem, and below 0.1% for the sixteen- and seventeen-class problems. The sinc methods appear to be comparable to the state-of-the-art in the four-class problem. For the sixteen- and seventeen-class subsets, the sinc methods surpass all published results.

6. CONCLUSIONS

We have shown that the SVM classification machine learning problem can be tackled in the context of signal theory. The interrelation between Paley-Wiener spaces and the sinc kernel has been exploited to form an explicit relationship between our information model and the sinc kernel hyper-parameter. By employing some recent work on sequency analysis, it has been shown that the nature of the model can be discerned. Consequently, a finite hyper-parameter search space was realised. Moreover, by introducing further assumptions, we have shown that the compromise between computational effort and search space sparseness can be managed sensibly. Finally, the approach achieves the best results so far on a much-studied AVIRIS data set.

7. REFERENCES

[1] M. Belkin and P. Niyogi, "Semi-supervised learning on Riemannian manifolds," *Machine Learning*, vol. 56, no. 1–3, pp. 209–239, 2004.

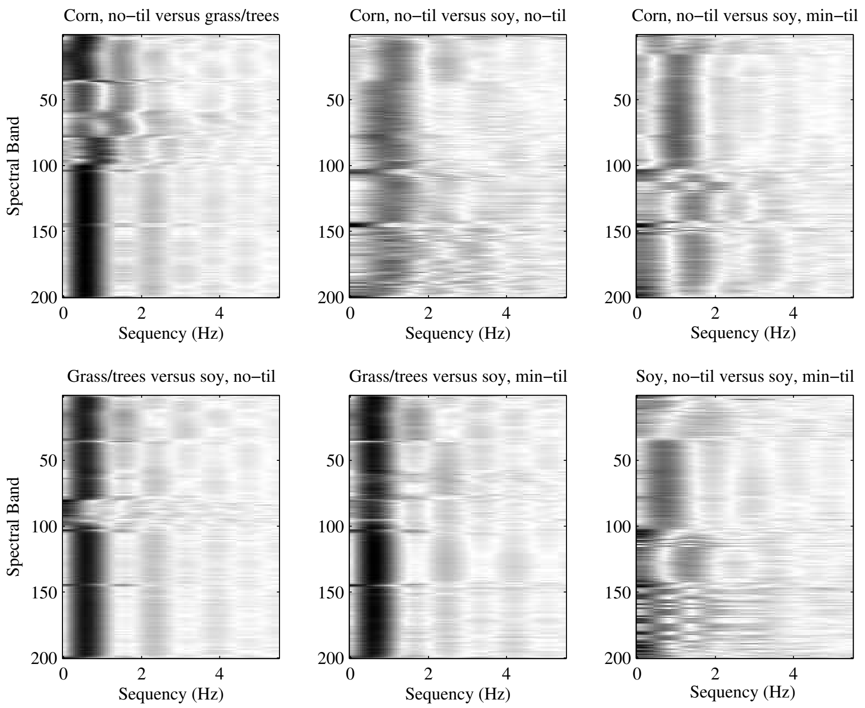

Fig. 2. Sequency spectra $|y^{\sim}|$ for the four-class AVIRIS problem. Darker tones indicate higher magnitude.

[2] G. Kimeldorf and G. Wahba, "A correspondence between Bayesian estimation of stochastic processes and smoothing by splines," *Annals of Mathematical Statistics*, vol. 41, no. 2, pp. 495–502, 1970.

[3] J. Mercer, "Functions of positive and negative type and their connection with the theory of integral equations," *Philosophical Transactions of Royal Society of London*, vol. 209, no. A456, pp. 415–446, 1909.

[4] A. J. Smola, B. Schölkopf, and K-R. Müller, "The connection between regularization operators and support vector kernels," *Neural Networks*, vol. 11, no. 4, pp. 637–649, 1998.

[5] J. D. B. Nelson, *The construction of some Riesz basis families and their application to coefficient quantization, sampling theory, and wavelet analysis*, PhD thesis, Anglia Polytechnic University, 2001.

[6] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *Journal of Machine Learning Research*, vol. 5 (Jan), pp. 27–72, 2004.

[7] R. Debnath and H. Takahashi, "Analyzing the behaviour of distribution of data in the feature space of SVM with Gaussian kernel," *Neural Information Processing Letters*, vol. 5, no. 3, pp. 41–48, 2004.

[8] B. Guo, R. Damper, S. Gunn, and J. Nelson, "Hyperspectral image fusion using spectrally weighted kernels," in *Proceedings of the 8th International Conference on Information Fusion*, Philadelphia, PA, 2005, paper B2-1, no pagination, Proceedings on CD-ROM.

[9] D. Landgrebe, "AVIRIS data," `ftp.ecn.purdue.edu`, last accessed 25/11/05.

[10] J. Gualtieri and R. Cromp, "Support vector machines for hyperspectral remote sensing classification," in *Proceeding of the 27th AIPR Workshop: Advances in Computer Assisted Recognition*, Washington DC, 1998, pp. 121–132.

[11] S. Tadjudin, *Classification of high dimensional data with limited training samples*, PhD thesis, School of Electrical Engineering and Computer Science, Purdue University, West Lafayette, IN, 1998.

[12] D. Landgrebe, "Hyperspectral image data analysis as a high dimensional signal processing problem," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 17–28, 2002.

[13] P. Du, "Self adaptive support vector machines and automatic feature selection," MSc thesis, McMaster University, Hamilton, Ontario, Canada, 2004.