Agnostic Learning versus Prior Knowledge in the Design of Kernel Machines

Gavin C. Cawley Nicola L. C. Talbot

***Abstract*— The optimal model parameters of a kernel machine are typically given by the solution of a convex optimisation problem with a single global optimum. Obtaining the best possible performance is therefore largely a matter of the *design* of a good kernel for the problem at hand, exploiting any underlying structure and optimisation of the regularisation and kernel parameters, i.e. model selection. Fortunately, analytic bounds on, or approximations to, the leave-one-out cross-validation error are often available, providing an efficient and generally reliable means to guide model selection. However, the degree to which the incorporation of *prior knowledge* improves performance over that which can be obtained using "standard" kernels with automated model selection (i.e. *agnostic learning*), is an open question. In this paper, we compare approaches using example solutions for all of the benchmark tasks on both tracks of the IJCNN-2007 Agnostic Learning versus Prior Knowledge Challenge.**

I. Kernel Learning Methods

Assume we are given labeled training data, $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{\ell}$, where $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ is a vector of input features describing the i^{th} example and $y_i \in \{-1, +1\}$ is an indicator variable such that $y_i = -1$ if the i^{th} example is drawn from class $\mathcal{C}^-$ and $y_i = +1$ is drawn from class $\mathcal{C}^+$. Kernel Ridge Regression [19] (or alternatively the Least-Squares Support Vector Machine [21]) aims to construct a linear model $f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}) + b$ in a fixed feature space, $\boldsymbol{\phi} : \mathcal{X} \rightarrow \mathcal{F}$, that is able to distinguish between examples drawn from $\mathcal{C}^-$ and $\mathcal{C}^+$, such that

$$\boldsymbol{x} \in \begin{cases} \mathcal{C}^+ & \text{if } f(\boldsymbol{x}) \geq 0 \\ \mathcal{C}^- & \text{otherwise} \end{cases} .$$

However, rather than specifying the feature space, $\mathcal{F}$ directly, it is implied by a kernel function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, giving the inner product between the images of vectors in the feature space, $\mathcal{F}$, i.e. $\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{\phi}(\boldsymbol{x}')$. A common kernel function is the Radial Basis Function (RBF) kernel

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \exp\left\{-\eta \|\boldsymbol{x} - \boldsymbol{x}'\|^2\right\}, \tag{1}$$

where η is a kernel parameter controlling the sensitivity of the kernel function. Other useful kernels include the Automatic Relevance Determination (ARD) kernel

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \exp\left\{-\sum_{i=1}^{d} \eta_i (x_i - x_i')^2\right\}, \tag{2}$$

which provides individual control over the sensitivity of the kernel to each of the input features, and the linear,

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x} \cdot \boldsymbol{x}' \tag{3}$$

Gavin Cawley and Nicola Talbot are with the School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, U.K.; E-mail: {gcc,nlct}@cmp.uea.ac.uk

and polynomial kernels

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x} \cdot \boldsymbol{x}' + c)^d \tag{4}$$

where c and d are kernel parameters ($d = 2$ gives the quadratic kernel and $d = 3$ the cubic kernel). The model parameters $(\boldsymbol{w}, b)$ are given by the minimum of a regularised [23] least-squares loss function,

$$\mathcal{L} = \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{1}{2\ell\mu}\sum_{i=1}^{\ell}\left[y_i - \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) - b\right]^2, \tag{5}$$

where μ is a regularisation parameter controlling the bias-variance trade-off [10]. The accuracy of the kernel machine on test data is critically dependent on the choice of good values for the *hyper-parameters*, in this case μ and η. The search for the optimal values for such hyper-parameters is a process known as *model selection*. The representer theorem [13] states that the solution to this optimisation problem can be written as an expansion of the form

$$\boldsymbol{w} = \sum_{i=1}^{\ell} \alpha_i \boldsymbol{\phi}(\boldsymbol{x}_i) \quad \Longrightarrow \quad f(\boldsymbol{x}) = \sum_{i=1}^{\ell} \alpha_i \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}) + b.$$

The dual parameters of the kernel machine are then given by the solution of a system of linear equations,

$$\begin{bmatrix} \boldsymbol{K} + \mu\ell\boldsymbol{I} & \boldsymbol{1} \\ \boldsymbol{1}^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix} = \begin{bmatrix} \boldsymbol{t} \\ 0 \end{bmatrix}. \tag{6}$$

which can be solved efficiently via Cholesky factorisation of $\boldsymbol{K} + \mu\ell\boldsymbol{I}$, with a computational complexity of $\mathcal{O}(\ell^3)$ operations [21].

A. Model Selection

An attractive feature of the kernel ridge regression machine is that it is possible to perform leave-one-out cross-validation [14, 16] in closed form, with minimal cost as a by-product of the training algorithm. Let $\boldsymbol{C}$ represent the matrix on the left hand side of (6), then the residual error for the i^{th} training pattern in the i^{th} fold of the leave-one-out process is given by,

$$r_i^{(-i)} = y_i - \hat{y}_i^{(-i)} = \frac{\alpha_i}{\boldsymbol{C}_{ii}^{-1}}. \tag{7}$$

Similar methods have been used in least-squares linear regression for many years, e.g [24]. While the optimal model parameters of the kernel machine are given by the solution of a simple system of linear equations, (6), some form of model selection is required to determine good values for the

hyper-parameters, $\boldsymbol{\theta} = (\mu, \boldsymbol{\eta})$ in order to maximise generalisation performance. The analytic leave-one-out cross-validation procedure described here can easily be adopted to form the basis of an efficient model selection strategy [6] based on a Allen's predicted residual sum-of-squares (PRESS) statistic [1],

$$\mathrm{PRESS}(\boldsymbol{\theta}) = \sum_{i=1}^{\ell} \left\{ r_i^{(-i)} \right\}^2 .$$

The PRESS criterion can be optimised efficiently using scaled conjugate gradient descent (e.g. [25]). For full details of the training and model selection procedures for the kernel ridge regression model, see [4]. The kernel machines used in this study were implemented using a MATLAB toolbox implementing a generalised form of kernel learning method, described in companion paper [5].

B. Performance Estimation

It not seem wise to be over-reliant on the validation set BER, available from the challenge website[1], to guide the development of models as it is far to small to provide a reliable indicator of the true level of generalisation performance, especially for highly imbalanced datasets, such as HIVA. A more reliable guide can be obtained via cross-validation [20] or bootstrap re-sampling [9] procedures using the labeled training set. For the previous Performance Prediction Challenge [12] and the Agnostic Learning track, we employed a computationally expensive, but reliable scheme based on 100-fold test-training splits of the available data. For the prior knowledge track, we adopt a more reasonable 10-fold cross-validation approach. It is important to avoid selection bias by performing model selection separately in each fold of the cross-validation procedure, i.e. we should view model selection as an integral part of the model fitting process.

II. Results Obtained on the ADA Dataset

The goal of the ADA benchmark is to identify high income individuals, earning $\$50$K per annum or more, on the basis of census data. The benchmark is derived from the Adult dataset from the UCI machine learning repository [17]. The data include a mixture of continuous, ordinal and Boolean features (e.g. age, education and sex respectively). This dataset seemed to present the least opportunity for incorporating prior knowledge into the design as of a kernel model as the pre-processing of the data for the agnostic track of the challenge is eminently sensible. We therefore followed the same pre-processing steps for both the agnostic and prior-knowledge submissions, with the exception of power transformations of the age, capital-gain and capital-loss continuous features, such that, e.g.

$$x_i^{\mathrm{age}} \leftarrow \sqrt[10]{x_i^{\mathrm{age}}}$$

This type of transformation [3] is commonly used to reduce the skew of the distribution of a feature having a heavy upper tail, making it better suited to distance-based kernel functions, such as the RBF kernel. Models were generated using both kernel ridge regression and kernel logistic regression, with a variety of kernel functions. Table I shows representative results for the ADA benchmark; the best results were obtained using kernel ridge regression, with an ARD kernel, which is currently the leading model on the prior knowledge track for this dataset, in terms of validation set performance. However, given that reliable cross-validation results are not yet available for this model, it would be unwise to *confidently* expect a similar level of performance on the test data.

TABLE I
REPRESENTATIVE RESULTS FOR THE ADA BENCHMARK.

model	kernel	cross-validation		validation set	
		BER	AUC	BER	AUC
KRR	linear	0.2004	0.8838	0.2206	0.8644
KRR	poly ($p = 2$)	0.1909	0.8948	0.2143	0.8745
KRR	poly ($p = 3$)	0.1920	0.8941	0.2094	0.8727
KRR	RBF	0.1949	0.8941	0.2095	0.8729
KRR	ARD	**0.1653**†	**0.9180**†	**0.1740**	**0.8910**

† biased leave-one-out estimate from the model selection process.

III. Results Obtained on the GINA Dataset

The GINA benchmark essentially describes an optical character recognition (OCR) problem, constructed from the MNIST[2] data [15], where the task is to distinguish the odd digits from the even. Each digit is represented by a grid of 28×28 integer pixel values in the range $[0\ 255]$, which we rescaled to lie in the range $[0\ 1]$, by dividing each feature by 255. For the agnostic track, the input vector consists of the pixel intensity values for two adjacent digits, the task being to determine whether the second digit is odd or even, so half of the input features represent uninformative *distractors*. The reference solutions for the agnostic track were implemented by training kernel ridge regression models with linear, quadratic, cubic and RBF kernels directly on the scaled input data. The results of performance estimation using 100 random training-test partitions of the data are shown in Table II. An improved agnostic solution, using an ARD kernel acting on the first hundred principal components of the data was later implemented, although external cross-validation proved prohibitively expensive, and so the performance estimate given here is the optimistically biased leave-one-out estimate used as the model selection criterion.

A. Engineered Solutions for the Prior Knowledge Track

The prior knowledge that the GINA dataset describes an optical character recognition problem, where each feature represents a pixel intensity on a regular grid, can be exploited in the design of the kernel. It seems reasonable to suggest that different areas of the grid are likely to carry

[1] http://www.agnostic.inf.ethz.ch/

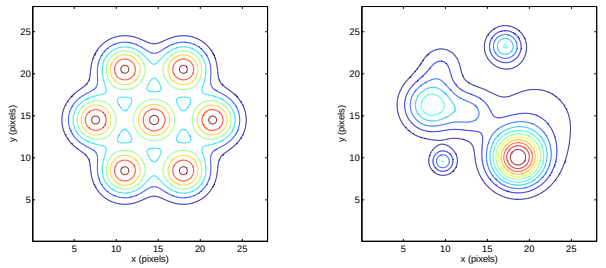
[2] http://yann.lecun.com/exdb/mnist

TABLE II

RESULTS ON THE GINA DATASET FOR THE AGNOSTIC LEARNING TRACK (BEST RESULTS SHOWN IN BOLD).

model	kernel	100-fold validation		validation set	
		BER	AUC	BER	AUC
KRR	linear	0.1324	0.9364	0.1273	0.9461
KRR	poly ($p = 2$)	0.0578	0.9848	0.0317	0.9940
KRR	poly ($p = 3$)	0.0532	0.9870	0.0285	0.9955
KRR	RBF	0.0571	0.9853	0.0442	0.9955
KRR	PCA-ARD	**0.0297**†	**0.9950**†	**0.0253**	**0.9968**

† biased leave-one-out estimate from the model selection process.



(a) (b)

Fig. 1. Initial distribution of receptive fields for the multiple receptive field (MRF) kernel (a) and the configuration following model selection (b).

more discriminative information than others, but that the variation in discriminative information is reasonably smooth across the image. The direct application of an ARD kernel would be computationally infeasible in this case as there are $28 \times 28 = 784$ hyper-parameters to be tuned (this would also be highly likely to result in over-fitting of the model selection criterion [7]). We therefore introduce the multiple receptive field (MRF) kernel, which is essentially an ARD kernel where the weights are given by the sum of seven Gaussian receptive fields distributed across the image. The twenty eight hyper-parameters of the MRF kernel describe the location, width and sensitivity of each of the receptive fields. Figure 1 (a) shows a contour plot of the initial weight matrix for the multiple receptive field RBF kernel. Through model selection, the hyper-parameters evolve so that the receptive fields focus on areas of the image containing the most discriminative information, as shown in Figure 1 (b).

The dataset for the prior knowledge tract also provides the identity of each digit comprising the training set. This is useful as the target concept is actually a composite of ten latent sub-classes, representing each individual digit. We therefore adopt a hierarchical approach, in which the first layer consists of 25 kernel ridge regression models, trained to distinguish between all possible pairs consisting of one odd digit and one even digit. The outputs of these kernel machines form the input to a kernel logistic regression model, used to estimate the *a-posteriori* probability that the input digit is odd. The design of the first layer networks was relatively straight-forward, with all networks being based on a simple radial basis function kernel, operating directly on the pixel intensity values. The regularisation and kernel parameters were optimised separately for each network, via minimisation of the PRESS statistic. While this may appear to be computationally rather expensive, the datasets used to train each network consisted of only those training patterns representing two of the ten digits. As the computational complexity of the training and model selection procedures are $\mathcal{O}(\ell^3)$, training the full set of low level networks is still approximately five time faster than training a model on the full dataset. The output classifier is trained on the leave-one-out output of the first layer networks, in order to provide a reasonably unbiased dataset that is more representative of operational conditions (c.f. [8]).

TABLE III

RESULTS ON THE GINA DATASET FOR THE PRIOR KNOWLEDGE TRACK (BEST RESULTS SHOWN IN BOLD).

model	kernel	cross validation		validation set	
		BER	AUC	BER	AUC
KRR	linear	0.1297	0.9416	0.1270	0.9525
KRR	poly ($p = 2$)	0.0365	0.9914	0.0158	0.9998
KRR	poly ($p = 3$)	0.0310	0.9938	0.0095	**0.9999**
KRR	poly ($p = 4$)	0.0284	0.9948	**0.0064**	**0.9999**
KRR	poly ($p = 5$)	0.0279	0.9949	**0.0064**	**0.9999**
KRR	poly ($p = 6$)	0.0256	0.9949	0.0126	**0.9999**
KRR	RBF	0.0290	0.9945	0.0095	0.9998
KRR	MRF	0.0315	0.9948	0.0157	0.9996
KRR+KRR	RBF+RBF	0.0263	0.9956	0.0128	0.9996
KRR+KRR	RBF+ARD	0.0253	0.9959	0.0192	0.9994
KRR+KRR	MRF+RBF	0.????†	0.????†	0.????	0.????
KRR+KRR	MRF+ARD	0.????†	0.????†	0.????	0.????

† biased leave-one-out estimate from the model selection process.

Table III shows example results for the Prior Knowledge track. In this case, we are able to significantly improve on the Agnostic Learning track entries, the best model is currently tied for first place on the Prior Knowledge track in terms of validation set BER. It seems likely that this is largely due to the deletion of the distractors (note that the best performance is still obtained using a relatively simple classifier). It is possible that the distractors we particularly malicious here, as they are highly correlated with each other, but describe a coherent, but uninformative structure within the data.

IV. RESULTS OBTAINED ON THE HIVA DATASET

The aim of the HIVA benchmark is to identify small molecules that are active against HIV based on their chemical structure. The Agnostic Track dataset provides a large set of binary molecular descriptors, computed using the ChemTK[3] package. The reference solutions for this dataset comprise of kernel ridge regression models with standard kernels acting directly on the binary features. In each case, the threshold,

[3] http://www.sageinformatics.com

regularisation and kernel parameters were optimised using the PRESS statistic. A similar approach was used to produce the first place entry for the corresponding benchmark in the WCCI-2006 Performance Prediction Challenge, with a test BER of 0.2757.

TABLE IV
RESULTS ON THE HIVA DATASET FOR THE AGNOSTIC LEARNING TRACK.

model	kernel	100-fold validation		validation set	
		BER	AUC	BER	AUC
KRR	linear	0.2547	0.8071	0.3311	0.6990
KRR	poly ($d = 2$)	**0.2444**	0.7991	0.2535	0.7253
KRR	poly ($d = 3$)	0.2523	0.8051	**0.2467**	0.7486
KRR	RBF	0.2495	**0.8092**	0.2819	**0.7604**

A. Engineered Solution for the Prior Knowledge Track

For the prior knowledge track, we obtained 1024 bit binary chemical fingerprints for each molecule, using the `generatemd` tool from the ChemAxon chemoinformatics suite[4]. These fingerprints, which represent structural properties of the molecule, are widely used in searching for similar molecules in large databases, or for screening molecules for putative pharmacological activity. These fingerprints provide a reasonable starting point for investigation of the HIVA benchmark. Work is currently ongoing to fine-tune these chemical fingerprints and to investigate other forms of structural descriptors. Representative results are shown in Table V; at the close of the development phase, the model based on a quadratic kernel is in first place, according to the validation set BER.

TABLE V
RESULTS ON THE HIVA DATASET FOR THE PRIOR KNOWLEDGE TRACK.

model	kernel	100-fold validation		validation set	
		BER	AUC	BER	AUC
KRR	linear	0.2957	**0.7988**	**0.2548**	0.7486
KRR	poly ($d = 2$)	0.2914	0.7411	0.2476	0.6786
KRR	poly ($d = 3$)	**0.2888**	0.7406	0.2629	**0.7741**
KRR	poly ($d = 4$)	0.2989	0.7365	0.3444	0.7384
KRR	RBF	0.4889	0.4573	0.5000	0.4519

A molecule can be viewed as a graph, with labeled vertices representing the atoms and weighted edges representing the chemical bonds. A walk through such a graph can then be represented as a string (e.g. `H-C-C=O`) giving the atoms visited and the strength of the bonds connecting them. A histogram, recording the counts of strings representing all possible walks of length k or less, then provides a sparse *molecular fingerprint* describing the structure of the molecule. It seems reasonable to suggest that similar molecules will share many common paths, and so a simple kernel function for small molecules simply computes the inner products between histograms [22]. This kernel can be computed efficiently using a *trie* or suffix tree structure [11]. Work is currently underway to investigate the use of such kernels for the HIVA dataset and on data integration, to assimilate kernels based on different sources of information.

[4] `http://www.chemaxon.com/`

V. RESULTS OBTAINED ON THE NOVA DATASET

The NOVA dataset consists of messages posted to various Usenet newsgroups, with messages posted to groups pertaining to religion or politics forming the positive class. For the Agnostic Learning track all words containing digits were removed and all letters converted to lower case. Short words with less than three letters were discarded, along with ≈ 2000 very common words. All words were then truncated to a maximum of seven letters. The input vector for each message then records the number of occurrences of each of 16,969 remaining distinct words comprising the corpus. Table VI shows representative results obtained by applying KRR models, with standard kernels and automated model selection to the standardized data. This simple approach appears to give highly competitive results, and the linear and cubic KRR classifiers have yet to be surpassed in terms of validation set BER.

TABLE VI
RESULTS ON THE NOVA DATASET FOR THE AGNOSTIC LEARNING TRACK.

model	kernel	100-fold validation		validation set	
		BER	AUC	BER	AUC
KRR	linear	**0.0491**	**0.9878**	**0.0440**	**0.9968**
KRR	poly ($d = 2$)	0.0550	0.9862	0.0640	0.9955
KRR	poly ($d = 3$)	0.0569	0.9854	**0.0044**	0.9947
KRR	RBF	0.0635	0.9828	0.0480	0.9942

A. Engineered Solution for the Prior Knowledge Track

The NOVA benchmark provides greater scope for engineering the data than many of the other benchmarks included in the challenge. Messages posted to newsgroups are often typed in haste and submitted without proof-reading. We therefore perform automated correction of mis-spellings as an optional stage in the pre-processing of the data, in order to improve the accuracy of term-matching. Many words vary only due to the presence of a suffix, which does not affect the information conveyed by the word. Stemming aims to strip redundant suffixes to obtain the *stem* or *root* of the word, e.g. reducing "fisher", "fishing" or "fished" to the stem "fish". Here we use the UEA-Lite stemmer[5]. Lastly, we adopt the term frequency-inverse document frequency (TF-IDF) coding scheme commonly used in text retrieval problems [18]. The

[5] http://www.cmp.uea.ac.uk/Research/stemmer

term frequency within a document is given by

$$tf = \frac{n_i}{\sum_k n_k},$$

where n_k records the number of occurrences of the k^{th} term. The inverse document frequency,

$$idf = \log\left\{\frac{|D|}{|d_k \supset t_i|}\right\}$$

provides a measure of the importance of a term, where $|D|$ represents the number of documents in the corpus and $|d_k \supset t_i|$ is the number of documents in which term t_i appears. Rather than using a simple count, we use $tf \cdot idf$, which has the effect of suppressing common terms, while amplifying rare, but informative terms. Table VII shows some preliminary results, note that we have been able to improve marginally on the performance of the equivalent models from the Agnostic Learning Track. Stemming appears to be helpful, providing the lowest validation set BER recorded so far, however the automated spell checking appears to have been too aggressive, and we are in the process of fine-tuning this element of the system.

TABLE VII
RESULTS ON THE NOVA DATASET FOR THE PRIOR KNOWLEDGE TRACK.

model	pre-processing	cross validation		validation set	
		BER	AUC	BER	AUC
KRR	none	**0.0432**	**0.9894**	0.0540	**0.9886**
KRR	stemming	0.0504	0.9890	**0.0360**	0.9878
KRR	spell+stem	0.0626	0.9817	0.0540	0.9782

Again as newsgroups are organised in an hierarchical manner, an alternative approach would involve the creation of a number of expert models, each of which is used to distinguish between different groups at the top level (e.g. `alt.*`-v-`comp.*`), an intermediate level (e.g. `comp.sys.*`-v-`talk.politics.*`) or the lowest level (e.g. `comp.graphics`-v-`talk.politics.misc`). These experts then provide the input features for a classifier used to identify messages posted to groups relating to religion or politics. This approach will be investigated at a later stage, although the performance of single classifiers, given appropriate pre-processing, is already good.

VI. RESULTS OBTAINED ON THE SYLVA DATASET

The SYLVA dataset describes the distribution of different tree species in four wilderness areas within the Roosevelt National Forest, located in northern Colorado, according to a set of cartographic variables [2], describing geographical location, terrain and soil type. While the original data are partitioned into classes representing seven different tree species (Spruce-Fir, Lodgepole Pine, Ponderosa Pine, Cottonwood/Willow, Aspen, Douglas-Fir and Krummholz), the aim of the SYLVA benchmark is to distinguish between Ponderosa Pine and all other species. For the Agnostic Learning track, the input vector is formed by the concatenation of cartographic features representing four patterns from the original dataset. Two of these patterns are used to decide the label and two are irrelevant (the positive class consisting of records where both key patterns represent Ponderosa Pine, the negative pattern consisting of records where neither represents Ponderosa Pine). This implies that half of the 108 input features are distractors and the remainder exhibit some degree of redundancy. Table VIII shows representative results for automated learning methods, using standard kernels (see [4] for further details).

TABLE VIII
RESULTS ON THE SYLVA DATASET FOR THE AGNOSTIC LEARNING TRACK.

model	kernel	100-fold validation		validation set	
		BER	AUC	BER	AUC
KRR	linear	0.0149	0.9982	0.0069	0.9980
KRR	poly ($d = 2$)	**0.0077**	**0.9991**	**0.0045**	0.0990
KRR	poly ($d = 3$)	0.0078	0.9990	**0.0045**	**0.9991**
KRR	RBF	0.0079	0.9990	0.0049	**0.9991**

A. Engineered Solution for the Prior Knowledge Track

For the Prior Knowledge track, the irrelevant input features are discarded, which should substantially reduce the difficulty of the task. The training set provides details of 26,172 distinct patterns from the original COVTYPE dataset. Table IX shows the distribution of cover type for each of the four wilderness areas, note that Ponderosa Pine are not found in the Rawah or Neota wilderness areas. In addition Ponderosa Pine are only found to exist in thirteen (1–6, 10, 11, 13, 14, 16, 17, and 32) of the forty soil types. We can therefore pre-classify any example containing a sub-pattern from Rawah or Neota or from any other soil type as belonging to the negative class. This leaves only 1,335 *difficult* patterns that must be classified. This is well within the reach of a kernel ridge regression model. A KRR model with an RBF kernel achieves a validation set BER of 0.0041 (joint 4$^{\text{th}}$ place) and an AUC of 0.9992, and performs slightly better than the corresponding Agnostic Track model.

VII. CONCLUSIONS

In this paper, we have presented solutions to both tracks of the IJCNN-07 Agnostic Learning versus Prior Knowledge

TABLE IX
COVER TYPE BY WILDERNESS AREA.

Cover Type	Rawah	Neota	Comanche Peak	Cache la Poudre
Spruce-Fir	4779	796	3919	0
Lodgepole Pine	6635	410	5609	135
Ponderosa Pine	0	0	663	947
Cottonwood/Willow	0	0	0	137
Aspen	174	0	245	0
Douglas-Fir	0	0	373	453
Krummholz	228	104	565	0
Total	11816	1310	11374	1672

Challenge. The reference solutions for the Agnostic Learning track rely on the use of a limited range of standard kernel functions and an automated model selection scheme to achieve a good level of generalisation performance. The prior knowledge solutions are currently placed first or joint first on four of the five benchmarks (ADA, GINA, HIVA and NOVA), third on SYLVA and also leading overall. However, these results are based on the validation set performance, which is known to have a high variance, and so the final standings may well be very different! It is interesting to see however that the provision of prior knowledge, suggesting solutions that exploit any hidden structure of the problem, or encouraging the use of bespoke kernels, has had relatively little effect on the results. The only dataset where the Prior Knowledge model performed appreciable better (GINA), and there the difference in performance seems due solely to the deletion of irrelevant input features, rather than the incorporation of prior knowledge, This suggests perhaps that automated model selection procedures are becoming a genuinely practical proposition (or alternatively just that we have not been sufficiently imaginative in applying our prior knowledge! ;-).

ACKNOWLEDGMENTS

The authors would like to thank the organisers of the Agnostic Learning versus Prior Knowledge Challenge, and the Data Representation Discovery Workshop as well as the WCCI-2006 and NIPS-2006 model selection workshops and challenges. We also thank the participants of the preceding workshops for interesting discussions that have had a significant impact on our views on model selection and performance estimation. Lastly, we thank Dan Smith (NOVA), Richard Harvey (GINA), Richard Morris (HIVA) and Lora Mak (HIVA) for their advice on data representation.

REFERENCES

[1] D. M. Allen. The relationship between variable selection and prediction. *Technometrics*, 16:125–127, 1974.

[2] J. A. Blackard and D. J. Dean. Comparative accuracies of artificial neural networks and discriminative analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, 24:131–151, 1999.

[3] G. E. P. Box and P. W. Tidwell. Transformation of the independent variables. *Technometrics*, 4(4):531–550, November 1962.

[4] G. C. Cawley. Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN-2006)*, pages 2970–2977, Vancouver, BC, Canada, July 16–21 2006.

[5] G. C. Cawley, G. J. Janacek, and N. L. C. Talbot. Generalised kernel machines. In *Proceedings of the IEEE/INNS International Joint Conference on Neural Networks (IJCNN-2007)* (submitted), Orlando, FL, USA, August12–17 2007.

[6] G. C. Cawley and N. L. C. Talbot. Gene selection in cancer classification using sparse logistic regression with Bayesian regularisation. *Bioinformatics*, 22(19):2348–2355, October 1 2006.

[7] G. C. Cawley and N. L. C. Talbot. Preventing over-fitting in model selection via Bayesian regularisation of the hyper-parameters. *Journal of Machine Learning Research* (accepted subject to minor revisions), 2007.

[8] G. C. Cawley, N. L. C. Talbot, R. J. Foxall, S. R. Dorling, and D. P. Mandic. Heteroscedastic kernel ridge regression. *Neurocomputing*, 57:105–124, March 2004.

[9] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Shapman & Hall, 1993.

[10] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilema. *Neural Computation*, 4(1):1–58, 1992.

[11] D. Gusfield. *Algorithms on strings, trees and sequences - computer science and computational biology*. Cambridge University Press, 1997.

[12] I. Guyon, A. R. S. A. Almadari, G. Dror, and J. M. Buhmann. Performance prediction challenge. In *Proceedings of the IEEE/INNS International Joint Conference on Neural Networks (IJCNN'06)*, pages 1649–1656, July 16-21 2006.

[13] G. S. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic.*, 33:82–95, 1971.

[14] P. A. Lachenbruch and M. R. Mickey. Estimation of error rates in discriminant analysis. *Technometrics*, 10(1):1–11, February 1968.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2424, November 1998.

[16] A. Luntz and V. Brailovsky. On estimation of characters obtained in statistical procedure of reognition (in Russian). *Technicheskaya Kibernetica*, 3, 1969.

[17] D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz. UCI repository of machine learning databases, University of California, Irvine, Department of Information and Computer Sciences, Irvine CA. http://www.ics.uci.edu/∼mlearn/MLRepository.html, 1998.

[18] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[19] C. Saunders, A. Gammermann, and V. Vovk. Ridge regression in dual variables. In J. Shavlik, editor, *Proceedings of the Fifteenth International Conference on Machine Learning (ICML-1998)*. Morgan Kaufmann, 1998.

[20] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, B*, 36(1):111–147, 1974.

[21] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vanderwalle. *Least Squares Support Vector Machines*. World Scientific Publishing, 2002.

[22] S. J. Swamidass, J. Chen, J. Bruand, P. Phung, L. Ralaivola, and P. Baldi. Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *Bioinformatics*, 21 Supplement 1:i359–i368, 2005.

[23] A. N. Tikhonov and V. Y. Arsenin. *Solutions of ill-posed problems*. John Wiley, New York, 1977.

[24] S. Weisberg. *Applied linear regression*. John Wiley and Sons, New York, second edition, 1985.

[25] P. M. Williams. A Marquardt algorithm for choosing the step-size in backpropagation learning with conjugate gradients. Cognitive Science Research Paper CSRP-229, University of Sussex, Brighton, U.K., February 1991.