Analysis of the IJCNN 2007 Agnostic Learning *vs.* Prior Knowledge Challenge

Isabelle Guyon
ClopiNet
Berkeley, CA 94708, USA
isabelle@clopinet.com

Amir Saffari
Graz University of Technology, Austria
amir@ymer.org

Gideon Dror
Academic College of Tel-Aviv-Yaffo, Israel
gideon@mta.ac.il

Gavin Cawley
University of East Anglia, UK
gcc@cmp.uea.ac.uk

Abstract

We organized a challenge for IJCNN 2007 to assess the added value of prior domain knowledge in machine learning. Most commercial data mining programs accept data pre-formatted in the form of a table, with each example being encoded as a linear feature vector. Is it worth spending time incorporating domain knowledge in feature construction or algorithm design or can off-the-shelf programs working directly on simple low-level features do better than skilled data analysts? To answer these questions, we formatted five datasets using two data representations. The participants to the "prior knowledge" track used the raw data, with full knowledge of the meaning of the data representation. Conversely, the participants to the "agnostic learning" track used a pre-formatted data table, with no knowledge of the identity of the features. The results indicate that black-box methods using relatively unsophisticated features work quite well and rapidly approach the best attainable performances. The winners on the prior knowledge track used feature extraction strategies yielding a large number of low-level features. Incorporating prior knowledge in the form of generic coding/smoothing methods to exploit regularities in data is beneficial, but incorporating actual domain knowledge in feature construction is very time consuming and seldom leads to significant improvements. The AL *vs.* PK challenge web site remains open for post-challenge submissions: `http://www.agnostic.inf.ethz.ch/` .

1 Introduction

There has been a lengthy philosophical and scientific debate as to whether or not the brains of children are a "tabula rasa", without prior knowledge of their environment. While it is still unclear whether or not this hypothesis holds for the neocortex, it cannot be refuted that specialized cortices connected to sensory inputs have evolved over millions of years to process information in a specialized manner. Hence, the brain benefits in a variety of learning tasks from advanced feature extraction that in some sense embodies a form of "prior knowledge". Such specialized pre-processing allows humans and animals to excel in tasks such as face recognition and speech segmentation. On the other hand, the brain is also capable of learning without the benefit of such specialized pre-processing, for instance, a human expert can learn to manage an investment portfolio, a task relying upon data representations not readily available from the sensory cortices. Perhaps evolution has led to an innate preference for "simple solutions" over different problems arising in a domain, allowing us to complete such tasks without a substantial amount of prior knowledge. Learning machines are tools designed to help engineers solve problems with as little expense in terms of human labor as possible. Incorporating "prior knowledge" or "domain knowledge" in a learning machine can be fairly intensive in labor and expertise, so researchers strive to improve their predictive models to provide good performance without substantial human intervention. In recent years, this has been made possible, even in cases where the number of examples is small compared to the dimension of the feature space, with the introduction of a new generation of regularized learning methods. For the purpose of this paper, we define "prior knowledge" as any form of knowledge about a given task that may be incorporated in the design of a predictive learning system, prior to training on the data. This may include: feature information (type of features, topological relationships between features, indications of feature relevance) and more general information about the nature and goal of the task that can reveal clusters in data, or the presence of

missing data, etc. For example, in a vision task in which images are encoded as gray level pixels, the knowledge of the nature of the features and their topological relationships allows the designer to perform specialized image filtering or to use specialized machine learning architectures, such as the convolutional neural network [17]. An alternative form of "domain knowledge" used in handwriting recognition is to model the dynamics of handwriting to extract relevant features from on-line data. Another example is the study of DNA or protein sequence data, where knowledge about the primary, secondary, and tertiary structure of a molecule improves the identification of active sites. This may be exploited in specialized kernels used with kernel machines [3].

We decided to assess the real added value of prior/domain knowledge in machine learning by organizing a competition, which we call the AL *vs.* PK (agnostic learning *vs.* prior knowledge) challenge. Challenges are important in several respects: Firstly they help us to answer questions of practical and/or scientific interest. Challenges are an inherently fair form of comparison as the test data are hidden from the participants and the aim of all the competitors is to win, and so the comparison involves learning methods applied with considerable care and effort, rather than "straw men". This enables the challenge to identify techniques that really work in practice. Secondly challenges push forward the state of the art and raise the standards of research, through the development of new methodologies. Lastly, they attract new researchers to the field and give them the opportunity to rapidly establish a good reputation within the research community. The challenge described in this paper was organized for IJCNN 2007. We received 1070 development entries from 50 groups and there were 13 ranked participants in each track for the final submission. In this paper we analyze the results of the challenge and present our findings.

2 Challenge Design

The design of the challenge was informed by experience gained from the previous competitions that we have organized [13, 14]. In particular, we used a system of on-line submission, which provided the competitors with immediate feed-back on a small subset of the data called validation set. The organizers provided initial submissions to bootstrap the challenge. A toolkit including some of the methods performing best in previous challenges was also provided (the so-called Challenge Learning Object Package CLOP [24]). At the end of a development period ending February 1^{st}, 2007, the validation set labels were revealed. The final ranking was performed on a large separate test set. The test set labels will remain hidden to permit meaningful comparison with post-challenge submissions.

The competition had two parallel tracks: For the "agnostic learning" (AL) track we supplied data preprocessed to provide a simple feature-based representation, suitable for use with any off-the-shelf machine learning or data mining package. The pre-processing used was identical to that used in the previous Performance Prediction Challenge, but with a new split of the data. The participants had no knowledge of the identity of the features in the agnostic track. New in this year's competition are the raw data representations used in the "prior knowledge" (PK) track, which are not necessarily in the form of data tables. For instance, in the drug discovery problem the raw data consists of a representation of the three dimensional structure of the drug molecules; in the text processing problem, the raw data are messages posted to USENET newsgroups. The participants had full knowledge of the meaning of the representation of the data in the PK track. Therefore, PK competitors had the opportunity to use domain knowledge to build better predictors and beat last year's AL results or make new "agnostic" entries. Note that the training/test splits used are the same in both tracks, but the example ordering is different in each data subset to hinder matching patterns in the two representations and/or submitting results with the representation prescribed for the other track.

The challenge started on October 1^{st}, 2006 and ended on August 1^{st}, 2007 (duration: 10 months). Two milestone rankings of the participants were made using the test set, without revealing either the test labels or the test performances: on December 1^{st}, for the model selection game, and on March 1^{st}, to allow us to publish intermediate results [15]. To be eligible for the final ranking, submissions had to include results on all the tasks of the challenge in either track, on the test data. However, recognizing that domain knowledge is task specific, prizes were given for each task individually in the "prior knowledge" track. For each group, only the last five entries in either track counted towards the final ranking.

We used the same five data sets as in the IJCNN 2006 challenge, but formatted differently. The tasks are five two-class classification problems spanning a variety of domains (marketing, handwriting recognition (HWR), drug

discovery, text classification, and ecology) and a variety of difficulties, with sufficiently many examples to obtain statistically significant results. The input variables are continuous or binary, sparse or dense. Some raw data representations are not feature based. In some problems, the class proportions are very imbalanced. A detailed report on the data preparation is available [12]. The main data characteristics are summarized in Table 1. Non-feature based representations are supplied for HIVA (molecular structure) and NOVA (emails) in the PK track.

Table 1: Datasets of the AL *vs.* PK challenge

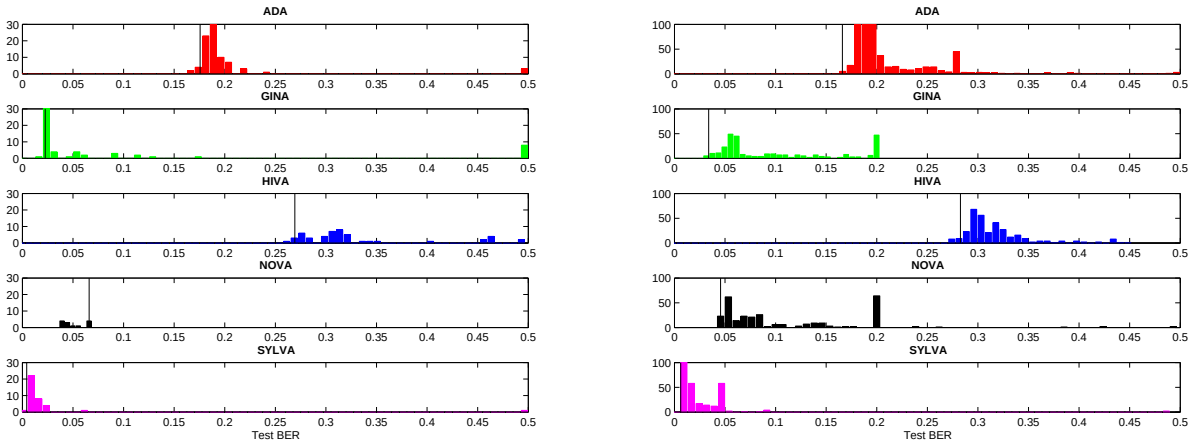
Dataset	Domain	Number of examples (training/validation/test)	Positive class (% num. ex.)	Number of features	
				Raw data (for PK)	Preprocessed (for AL)
ADA	Marketing	4147 / 415 / 41471	28.4	14	48
GINA	HWR	3153 / 315 / 31532	49.2	784	970
HIVA	Drug discovery	3845 / 384 / 38449	3.5	Molecules	1617
NOVA	Text classification	1754 / 175 / 17537	28.5	Text	16969
SYLVA	Ecology	13086 / 1309 / 130857	6.2	108	216

Table 2: PK better than AL comparison results

	ADA	GINA	HIVA	NOVA	SYLVA
Min PK BER	0.170	**0.019**	**0.264**	**0.037**	**0.004**
Min AL BER	**0.166**	0.033	0.271	0.046	0.006
Median PK BER	**0.189**	**0.025**	0.310	**0.047**	**0.008**
Median AL BER	0.195	0.066	**0.306**	0.081	0.015
Pval ranksum test	$\mathbf{5\ 10^{-8}}$	$\mathbf{3\ 10^{-18}}$	0.25	$\mathbf{8\ 10^{-6}}$	$\mathbf{10^{-18}}$
Jorge Sueiras				−	
Juha Reunanen [22]		+			+
Marc Boullé [7]	+	+		−	-
Roman Lutz [18]					+
Vladimir Nikulin [20]	−	+			+
Vojtech Franc	+	+			
CWW		−	−		
Reference (gcc) [8]	+	+	−		
Pvalue sign test	0.31	0.19	0.25	0.25	0.31

3 Results of the AL *vs.* PK challenge

The final ranking of submissions was based on the balanced error rate (BER) on the test set. The BER is the average of the error rate on the positive class and the error rate of the negative class. The Area Under the *ROC* Curve (AUC) was also computed, but not used for scoring. To obtain the overall ranking we averaged the ranks of participants in each track after normalizing by the number of entries. The number of submissions was unlimited, but only the five last "complete" submissions for each entrant in either track were included in the final ranking. For the first few weeks of the challenge, the top of the rankings were largely dominated by agnostic track (AL) submissions. However, the learning curves for the agnostic learning and prior knowledge tracks eventually crossed for all datasets, except for ADA. After approximately 150 days the PK performance asymptote was reached. The asymptotic performances are reported at the top of Table 2. In contrast, in the IJCNN-06 performance prediction challenge, using the same data as the AL track, the competitors attained almost their best performances within about 60 days and kept improving only slightly afterward.



(a) (b)

Figure 1: Distribution of test set Balanced Error Rate (BER). (a) Prior knowledge (PK) track. (b) Agnostic learning (AL) track. The thin vertical line indicates the best ranked entry (only the 5 last of each participant are ranked).

Figure 1, shows the distribution of the test BER for all entries. There were approximately 60% more submissions for the AL track than in the PK track. This indicates that the "prior knowledge" track was harder to enter. However, **the participants who did enter the PK track performed significantly better on average than those who entered the AL track**, on all datasets except for HIVA. To quantify this observation we ran a Wilcoxon rank sum test on the difference between the median values of the two tracks (Table 2). We also performed paired comparisons for entrants who entered both tracks, using their last 5 submissions. In Table 2, a "+" indicates that the entrant performed best in the PK track and a "-" indicates the opposite. We see that **the entrants who entered both tracks did not always succeed in obtaining better results in the PK track**. The p-values of the sign test do not reveal a significant dominance of PK over AL or vice versa in that respect (all are between 0.25 and 0.5). However, for HIVA and NOVA the participants who entered both tracks failed to get better results in the PK track. We conclude that, while on average PK seems to win over AL, success is uneven and depends both on the domain and on the individuals' expertise.

Agnostic learning methods

The winner of the "agnostic learning" track is Roman Lutz, who also won the Performance Prediction Challenge (IJCNN06) [18], using boosting techniques. Gavin Cawley, who joined the organization team and was co-winner of the previous challenge, made a reference entry using LSSVMs, which slightly outperforms that of Lutz. The improvements he made can partly be attributed to the introduction of an ARD kernel, which automatically down-weighs the least relevant features and to a Bayesian regularization at the second level of inference [8, 9]. The second best entrant is the Intel group, also using boosting methods. The next best ranking entrants include Juha Reunanen and Hugo Jair Escalante, who have both been using CLOP models provided by the organizers and have proposed innovative search strategies for model selection: Escalante is using a biologically inspired particle swarm technique [11, 10] and Reunanen a cross-indexing method to make cross-validation more computationally efficient [22, 23]. Other top ranking participants in the AL track include Vladimir Nikulin [20] and Jörg Wichard [27] who both experimented with several ensemble methods, Erinija Pranckeviciene [21] who performed a study of linear programming SVM methods, and Marc Boullé who introduced a new data grid method [6]. In the following sections, we look into more details at the methods employed in the "prior knowledge" track to outperform the results of the "agnostic track".

ADA: the marketing application

The task of ADA is to discover high revenue people from census data, presented in the form of a two-class classification problem. The raw data from the census bureau is known as the Adult database in the UCI machine-learning repository [16]. The 14 original attributes (features) represent age, workclass, education, education, marital status, occupation, native country, etc. and include continuous, binary and categorical features. The PK track had access to the original features and their descriptions. The AL track had access to a preprocessed numeric representation of the features, with a simple disjunctive coding of categorical variables, but the identity of the features was not revealed. We expected that the participants of the AL *vs.* PK challenge could gain in performance by optimizing the coding of the input features. Strategies adopted by the participants included using a thermometer code for ordinal variables (Gavin Cawley) and optimally grouping values for categorical variables (Marc Boullé). Boullé also optimally discretized continuous variables, which make them suitable for a naïve Bayes classifier. However, the advantage of using prior knowledge for ADA was marginal. The overall winner on ADA is in the agnostic track (Roman Lutz), and the entrants who entered both tracks and performed better using prior knowledge do not have results statistically significantly better. We conclude that optimally coding the variables may be not so crucial and that good performance can be obtained with a simple coding and a state-of-the-art classifier.

GINA: the handwriting recognition application

The task of GINA is handwritten digit recognition, the raw data is known as the MNIST dataset [17]. For the "agnostic learning" track we chose the problem of separating two-digit odd numbers from two-digit even numbers. Only the unit digit is informative for this task, therefore at least 1/2 of the features are distractors. Additionally, the pixels that are almost always blank were removed and the pixel order was randomized to hide the meaning of the features. For the "prior knowledge" track, only the informative digit was provided in the original pixel map representation. In the PK track the identities of the digits (0 to 9) were provided for training, in addition to the binary target values (odd *vs.* even number). Since the prior knowledge track data consists of pixel maps, we expected the participants to perform image pre-processing steps such as noise filtering, smoothing, de-skewing, and feature extraction (points, loops, corners) and/or use kernels or architectures exploiting geometric invariance by small translation, rotation, and other affine transformations, which have proved to work well on this dataset [17]. Yet, the participants to the PK track adopted very simple strategies, not involving a lot of domain knowledge. Some just relied on the performance boost obtained by the removal of the distractor features (Vladimir Nikulin, Marc Boullé, Juha Reunanen). Others exploited the knowledge of the individual class labels and created multi-class of hierarchical classifiers (Vojtech Franc, Gavin Cawley). Only the reference entries of Gavin Cawley (which obtained the best BER of 0.0192) included domain knowledge by using an RBF kernels with tunable receptive fields to smooth the pixel maps. In the future, it would be interesting to assess the methods of Simard et al [25] on this data to see whether further improvements are obtained by exploiting geometrical invariances. The agnostic track data was significantly harder to analyze because of the hidden class heterogeneity and the presence of feature distractors. The best GINA final entry was therefore on the PK track and all four ranked entrants who entered both tracks obtained better results in the PK track. Further, the differences in performance are all statistically significant.

HIVA: the drug discovery application

The task of HIVA is to predict which compounds are active against the AIDS HIV infection. The original data from the NCI [1] has 3 classes (active, moderately active, and inactive). We brought it back to a two-class classification problem (active & moderately active *vs.* inactive), but we provided the original labels for the "prior knowledge" track. The compounds are represented by their 3d molecular structure for the "prior knowledge" track (in SD format). For the "agnostic track" we represented the input data as vector of 2000 sparse binary variables. The variables represent properties of the molecule inferred from its structure by the ChemTK software package (version 4.1.1, Sage Informatics LLC). The problem is therefore to relate structure to activity (a QSAR - quantitative structure-activity relationship problem) to screen new compounds before actually testing them (a HTS - high-throughput screening problem). Note

that in such applications the BER is not the best metric to assess performances since the real goal is to identify correctly the compounds most likely to be effective (belonging to the positive class). We resorted to using the BER to make comparisons easier across datasets. The raw data was not supplied in a convenient feature representation, which made it impossible to enter the PK track using agnostic learning methods, using off-the-shelf machine learning packages. The winner in HIVA (Chloé-Agathe Azencott of the Pierre Baldi Laboratory at UCI) is a specialist in this kind of dataset, on which she is working towards her PhD [2]. She devised her own set of low level features, yielding a "molecular fingerprint" representation, which outperformed the ChemTK features used on the agnostic track. Her winning entry has a test BER of 0.2693, which is significantly better than the test BER of the best ranked AL entry of 0.2827 (error bar 0.0068). The results on HIVA are quite interesting because most agnostic learning entrants did not even attempt to enter the prior knowledge track and the entrants that did submit models for both tracks failed to obtain better results in the PK track. One of them working in an institute of pharmacology reported that too much domain knowledge is sometimes detrimental; experts in his institute advised against using molecular fingerprints, which ended up as the winning technique.

NOVA: the text classification application

The data of NOVA come from the 20-Newsgroup dataset [19]. Each text to classify represents a message that was posted to one or several USENET newsgroups. The raw data is provided in the form of text files for the "prior knowledge" track. The preprocessed data for the "agnostic learning" track is a sparse binary representation using a bag-of-words with a vocabulary of approximately 17000 words (the features are simply frequencies of words in text). The original task is a 20-class classification problem but we grouped the classes into two categories (politics and religion *vs.* others) to make it a two-class problem. The original class labels were available for training in the PK track but not in the AL track. As the raw data consist of texts of variable length it was not possible to enter the PK track for NOVA without performing a significant pre-processing. All PK entrants in the NOVA track used a bag-of-words representation, similar to the one provided in the agnostic track. Standard tricks were used, including stemming. Gavin Cawley used the additional idea of correcting the emails with an automated spell checker. No entrant who entered both tracks outperformed their AL entry with their PK entry in their last ranked entries, including the winner! This is interesting because the best PK entries made throughout the challenge significantly outperform the best AL entries (BER difference of 0.0089 for an error bar of 0.0018), see also Figure 1. Hence in this case, **the PK entrants overfitted and were unable to select among their PK entries those, which would perform best on test data**. This is not so surprising because the validation set on NOVA is quite small (175 examples). Even though the bag-of-words representation is known to be state-of-the-art for this kind of applications, it would be interesting to compare it with more sophisticated representations. To our knowledge, the best results on the 20 Newsgroup data were obtained by the method of distributional clustering by Ron Bekkerman [4].

SYLVA: the ecology application

The task of SYLVA is to classify forest cover types. The forest cover type for 30 x 30 meter cells was obtained from US Forest Service (USFS) Region 2 Resource Information System (RIS) data [5]. We converted this into a two-class classification problem (classifying Ponderosa pine *vs.* everything else). The input vector for the "agnostic learning" track consists of 216 input variables. Each pattern is composed of 4 records: 2 true records matching the target and 2 records picked at random. Thus 1/2 of the features are distractors. The "prior knowledge" track data is identical to the "agnostic learning" track data, except that the distractors are removed and the meaning of the features is revealed. For that track, the identifiers in the original forest cover dataset are revealed for the training set. As the raw data was already in a feature vector representation, this task was essentially testing the ability of the participants in the AL track to perform well in the presence of distractor features. The PK track winner (Roman Lutz) in his Doubleboost algorithm exploited the fact that each pattern was made of two records of the same pattern to train a classifier with twice as many training examples. Specifically, a new dataset was constructed by putting the second half of the data (variables 55 to 108) below the first half (variables 1 to 54). The new dataset is of dimension 2n times 54 (instead of n times 108). This new dataset is used for fitting the base learner (tree) of his boosting algorithm. The output

of the base learner is averaged over the two records belonging to the same pattern. This strategy can be related to the neural network architectures using "shared weights", whereby at training time, the weights trained on parts of the pattern having similar properties are constrained to be identical [17]. This reduced the number of free parameters of the classifier.

4 Conclusion

This paper presented the results of the IJCNN 2007 competition, whose goal was to compare two approaches in machine learning: the "agnostic learning" (AL) approach putting all the effort on the classifier and the "prior knowledge" (PK) approach capitalizing on human domain knowledge. The challenge was very successful in attracting a large number of participants who competed in the two tracks. For the first few months of the challenge, AL lead over PK, showing that the development of good AL classifiers is considerably faster. As of March 1^{st} 2007, PK was leading over AL on four out of five datasets. We extended the challenge five more months, but few significant improvements were made during that time period. On datasets not requiring real expert domain knowledge (ADA, GINA, SYLVA), the participants entering both track obtained better results in the PK track, using a special-purpose coding of the inputs and/or the outputs, exploiting the knowledge of which features were uninformative, and using "shared weights" for redundant features. On the datasets requiring most real expert domain knowledge (HIVA and NOVA), several entrants failed to capitalize on prior knowledge. For both HIVA and NOVA, the winning data representation consisted of a high-dimensional vector of low level features ("molecular fingerprints" and "bag-of-words"). From the analysis of this challenge, we conclude that agnostic learning methods are very powerful. They quickly yield (in 40 to 60 days) a level of performance close to the best achievable performance. General-purpose techniques for exploiting prior knowledge in the encoding of inputs or outputs or the design of the learning machine architecture (*e.g.* via shared weights) may provide an additional performance boost, but exploiting real domain knowledge is both difficult and time consuming. This fact seems to be a recurrent theme in machine learning publications and further confirmation is provided by the results of our challenge. Future work includes incorporating the best identified methods in our challenge toolkit, CLOP. The challenge web site remains open for post-challenge submissions at `http://www.agnostic.inf.ethz.ch/`, where supplementary analyzes and complete result tables are also made available.

Acknowledgments

We are very thankful to the institutions that originally provided the data. The organization of this challenge was a team effort to which many have participated. We are particularly grateful to Olivier Guyon (MisterP.net) our webmaster. Prof. Joachim Buhmann (ETH Zurich) who provided computer resources and all the advisors, beta-testers and sponsors are gratefully acknowledged (see `http://www.agnostic.inf.ethz.ch/credits.php` for a full list). The Challenge Learning Object Package (CLOP) is based on code to which many people have contributed [26, 24]. This project is supported by the Pascal network of excellence funded by the European Commission and the National Science Foundation under Grants N0. ECCS-0424142 and N0. ECCS-0736687. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Amir Safari acknowledges the support of the FWF Austrian Joint Research Project Cognitive Vision under projects S9103-N04 and S9104-N04 and the EU FP6-507752 NoE MUSCLE IST project.

References

[1] J. M. Collins Associate Director. The DTP AIDS antiviral screen program. `http://dtp.nci.nih.gov/docs/aids/aids_data.html`, 1999.

[2] C. A. Azencott, A. Ksikes, S. J. Swamidass, J. H. Chen, L. Ralaivola, and P. Baldi. One- to four-dimensional kernels for virtual screening and the prediction of physical, chemical, and biological properties. *J. Chem. Inf. Model.*, 2007 `http://pubs3.acs.org/acs/journals/doilookup?in_doi=10.1021/ci600397p`.

[3] P. Baldi. *Bioinformatics: The machine learning approach*. The MIT press, Cambridge, Massachusetts, 2001.

[4] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter. Distributional word clusters vs words for text categorization. 2003. Code available at `http://www.cs.technion.ac.il/~ronb/`.

[5] J. A. Blackard and D. J. Dean. Forest cover type. `http://kdd.ics.uci.edu/databases/covertype/covertype.html`, 1998.

[6] M. Boullé. Compression-based averaging of selective naive bayes classifiers. *JMLR*, pages 1659–1685, July 2007.

[7] M. Boullé. Report on preliminary experiments with data grid models in the agnostic learning vs. prior knowledge challenge. In *Proc. IJCNN07*, Orlando, Florida, Aug 2007. INNS/IEEE.

[8] G. C. Cawley and N. L. C. Talbot. Agnostic learning versus prior knowledge in the design of kernel machines. In *Proc. IJCNN07*, Orlando, Florida, Aug 2007. INNS/IEEE.

[9] G. C. Cawley and N. L. C. Talbot. Preventing over-fitting during model selection using Bayesian regularisation. *JMLR*, 8:841–861, April 2007.

[10] H. J. Escalante. Particle swarm model selection. *Submitted to JMLR*, 2007.

[11] H. J. Escalante, M. Montes, and L. E. Sucar. PSMS for neural networks: Results on the IJCNN 2007 agnostic *vs.* prior knowledge challenge. In *Proc. IJCNN07*, Orlando, Florida, Aug 2007. INNS/IEEE.

[12] I. Guyon. Datasets for the agnostic learning *vs.* prior knowledge competition. Technical report, Clopinet, 2005, `http://clopinet.com/isabelle/Projects/agnostic/Dataset.pdf`.

[13] I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror. Result analysis of the nips 2003 feature selection challenge. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 545–552. MIT Press, Cambridge, MA, 2005, `http://books.nips.cc/papers/files/nips17/NIPS2004_0194.pdf`.

[14] I. Guyon, A. Saffari, G. Dror, and J. Buhmann. Performance prediction challenge. In *IEEE/INNS conference IJCNN 2006*, Vancouver, July 16-21 2006.

[15] I. Guyon, A. Saffari, G. Dror, and G. Cawley. Agnostic *vs.* prior knowledge challenge. In *Proc. IJCNN07*, Orlando, Florida, Aug 2007. INNS/IEEE.

[16] R. Kohavi and B. Becker. The Adult database. `ftp://ftp.ics.uci.edu/pub/machine-learning-databases/adult/`, 1994.

[17] Y. LeCun and C. Cortes. The MNIST database of handwritten digits. `http://yann.lecun.com/exdb/mnist/`, 1998.

[18] R. W. Lutz. Logitboost with trees applied to the WCCI 2006 performance prediction challenge datasets. In *Proc. IJCNN06*, pages 2966–2969, Vancouver, Canada, July 2006, `http://stat.ethz.ch/~lutz/publ/WCCIlogitboost.php`. INNS/IEEE.

[19] T. Mitchell. The 20 Newsgroup dataset. `http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html`, 1999.

[20] V. Nikulin. Non-voting classification with random sets and boosting. In *Proc. IJCNN07*, Orlando, Florida, Aug 2007. INNS/IEEE.

[21] E. Pranckeviciene, R. Somorjai, and M. N. Tran. Feature/model selection by the linear programming SVM combined with state-of-art classifiers: What can we learn about the data. In *Proc. IJCNN07*, Orlando, Florida, Aug 2007. INNS/IEEE.

[22] J. Reunanen. Model selection and assessment using cross-indexing. In *Proc. IJCNN07*, Orlando, Florida, Aug 2007. INNS/IEEE.

[23] J. Reunanen. Resubstitution error is useful for guiding feature selection. *Submitted to JMLR*, 2007.

[24] A. Saffari and I. Guyon. Quick start guide for CLOP. Technical report, Graz University of Technology and Clopinet, May 2006. `http://ymer.org/research/files/clop/QuickStartV1.0.pdf`.

[25] P. Simard, D. Steinkraus, and J. Platt. Best practice for convolutional neural networks applied to visual document analysis. In *International Conference on Document Analysis and Recogntion (ICDAR)*, pages 958–962, Los Alamitos, 2003. IEEE Computer Society.

[26] J. Weston, A. Elisseeff, G. Bakir, and F. Sinz. The Spider machine learning toolbox. *`http://www.kyb.tuebingen.mpg.de/bs/people/spider/`*, 2005.

[27] J. Wichard. Agnostic learning with ensembles of classifiers. In *Proc. IJCNN07*, Orlando, Florida, Aug 2007. INNS/IEEE.