JMLR: Workshop and Conference Proceedings 16 (2011) 47–57 Workshop on Active Learning and Experimental Design

Baseline Methods for Active Learning

Gavin C. Cawley GCC@CMP.UEA.AC.UK
School of Computing Sciences
University of East Anglia
Norwich, Norfolk, NR4 7TJ, United Kingdom

Editor: I. Guyon, G. Cawley, G. Dror, V. Lemaire, and A. Statnikov

Abstract

In many potential applications of machine learning, unlabelled data are abundantly available at low cost, but there is a paucity of labelled data, and labeling unlabelled examples is expensive and/or time-consuming. This motivates the development of active learning methods, that seek to direct the collection of labelled examples such that the greatest performance gains can be achieved using the smallest quantity of labelled data. In this paper, we describe some simple pool-based active learning strategies, based on optimally regularised linear [kernel] ridge regression, providing a set of baseline submissions for the Active Learning Challenge. A simple random strategy, where unlabelled patterns are submitted to the oracle purely at random, is found to be surprisingly effective, being competitive with more complex approaches.

Keywords: pool based active learning, ridge regression

1. Introduction

The rapid development of digital storage devices has led to ever increasing rates of data capture in a variety of application domains, including text processing, remote-sensing, astronomy, chemoinformatics and marketing. In many cases the rate of data capture far exceeds the rate at which data can be manually labelled for the use of traditional supervised machine learning methods. As a result, large quantities of unlabelled data are often available at little or no cost, but obtaining more than a comparatively small amount of labelled data is prohibitively expensive or time consuming. Active learning aims to address this problem by constructing algorithms that are able to guide the labeling of a small amount of data, such that the generalisation ability of the classifier is maximized whilst minimising the use of the oracle. In pool-based active learning, a large number of unlabelled examples are provided from the outset, and training proceeds iteratively. At each step the active learning strategy chooses one or more unlabelled patterns to submit to the oracle, and the classifier updated using the newly acquired label(s). Pool-based active learning is appropriate in many applications, for instance drug design, where the aim is to predict the activity of a molecule against a virus, such as HIV, based on chemometric descriptors. A large number of small molecules have been subjected to chemometric analysis providing a large library of unlabelled data, however *in-vitro* testing is expensive. Active learning would therefore be useful in reducing the cost of drug design by targeting the effort *in-vitro* testing only on those molecules likely to be effective. There is a significant overlap between

© 2011 G.C. Cawley.

active learning and unsupervised or semi-supervised learning as the need for labelled data may be minimised by a learning algorithm that is able to take advantage of the information contained in the unlabelled examples. For a more detailed overview of active learning, see Settles (2009).

This paper describes a set of simple baseline solutions for an open challenge in active learning, described in detail in Guyon et al. (2010). The remainder of the paper is structured as follows: Section 2 provides a brief technical description of the base classifier and active learning strategies employed. Section 3 presents the results obtained using the baseline methods for the development and test benchmark datasets. Finally the work is summarised and conclusions presented in Section 4.

2. Technical Description of Baseline Methods

This section describes the technical detail of the baseline submissions, based on optimally regularised ridge regression, with the pre-processing steps employed, and three very simple active learning strategies.

2.1. Optimally Regularised [Kernel] Ridge Regression

Linear ridge regression is used as the base classifier for those baseline methods for the active learning challenge described in this paper. While more complex non-linear methods could have been used, such as a decision tree (Quinlan, 1986), support vector machine (Boser et al., 1992; Cortes and Vapnik, 1995) or naïve Bayes (e.g. Webb, 2002) classifier , very little labelled data is available at the start of the active learning process, and so a more complex classifier would run a greater risk of over-fitting. In addition, these methods were intended to provide a reasonably competitive baseline representing a fairly basic approach to the problem, and so a simple linear classifier seemed most appropriate. Let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{\ell}$ represent the training sample, where $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ is a vector of explanatory features for the i^{th} sample, and $y_i \in \{+1, \;-1\}$ is the corresponding response indicating whether the sample belongs to the positive or negative class respectively. Ridge regression provides a simple and effective classifier that is equivalent to a form of regularised linear discriminant analysis. The output of the ridge regression classifier, $\hat{y}_i$, and vector of model parameters, $\boldsymbol{\beta} \in \mathbb{R}^d$, are given by

$$\hat{y}_i = \boldsymbol{x}_i \cdot \boldsymbol{\beta} \qquad \text{and} \qquad \left[\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}\right]\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{y}, \tag{1}$$

where $\boldsymbol{X} = [\boldsymbol{x}_i]_{i=1}^{\ell}$ is the data matrix, $\boldsymbol{y} = (y_i)_{i=1}^{\ell}$ is the response vector and the ridge parameter, λ, controls the bias-variance trade-off (Geman et al., 1992). Note that classifiers used throughout this study included an unregularised bias parameter, which has been neglected here for notational convenience. Careful tuning of the ridge parameter allows the ridge regression classifier to be used even in situations with many more features than training patterns (i.e. $d \gg \ell$) without significant over-fitting (e.g. Cawley, 2006). Fortunately the ridge parameter can be optimised efficiently by minimising a closed-form leave-one-out cross-validation estimate of the sum of squared errors, i.e. Allen's PRESS statistic (Allen, 1974),

$$P(\lambda) = \frac{1}{\ell}\sum_{i=1}^{\ell}\left[\hat{y}_i^{(-i)} - y_i\right]^2 \qquad \text{where} \qquad \hat{y}_i^{(-i)} - y_i = \frac{\hat{y}_i - y_i}{1 - h_{ii}}, \tag{2}$$

$\hat{y}_i^{(-i)}$ represents the output of the classifier for the i^{th} training pattern in the i^{th} fold of the leave-one-out procedure and h_{ii} is an element of the principal diagonal of the hat matrix $\boldsymbol{H} = \boldsymbol{X}\left[\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}\right]^{-1}\boldsymbol{X}^T$. The ridge parameter can be optimised more efficiently in canonical form (Weisberg, 1985) via eigen-decomposition of the data covariance matrix $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{V}^T\boldsymbol{\Lambda}\boldsymbol{V}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix containing the eigenvalues. The normal equations and hat matrix can then be written as

$$[\boldsymbol{\Lambda} + \lambda\boldsymbol{I}]\,\boldsymbol{\alpha} = \boldsymbol{V}^T\boldsymbol{X}^T\boldsymbol{y} \quad \text{where} \quad \boldsymbol{\alpha} = \boldsymbol{V}^T\boldsymbol{\beta} \qquad \text{and} \qquad \boldsymbol{H} = \boldsymbol{V}\,[\boldsymbol{\Lambda} + \lambda\boldsymbol{I}]^{-1}\,\boldsymbol{V}^T \tag{3}$$

As only a diagonal rather than a full matrix need now be inverted following a change in λ, the computational expense of optimising the ridge parameter is greatly reduced. For problems with more features than training patterns, $d > \ell$, the kernel ridge regression classifier (Saunders et al., 1998) with a linear kernel is more efficient and exactly equivalent. The ridge parameter for KRR can also be optimised efficiently via an eigen-decomposition of the kernel matrix (Saadi et al., 2007).

2.2. Pre-processing

The following pre-processing steps were used for all datasets: First all constant features are deleted, including features where all values are missing. Binary fields are coded using the values 0 and 1. Categorical and ordinal variables are encoded using a 1-of-n representation, where n is the number of discrete categories/values. Missing values are imputed using the arithmetic mean, and dummy variables are added to indicate the pattern of missing data for each feature. Lastly, continuous features are transformed to have a standard normal distribution, by evaluating the inverse standard normal cumulative distribution function for the normalised rank for each observation. It is hoped that this transformation prevents variables with highly skewed distributions from having a disproportionate effect on the classifier, whilst still allowing the extreme values to lie in identifiable ails of the distribution.

2.3. Pool Based Active Learning

A number of very basic strategies for pool based active learning, suitable for use as baseline submissions, are easily identified:

- **Passive Learning:** All patterns submitted to the oracle for labeling in the first step. This is not strictly speaking an active learning strategy, but it provides a useful baseline for comparison.
- **Random sampling:** At each iteration, one or more unlabelled samples are selected at random to be labelled by the oracle. This is perhaps the most basic algorithm for pool-based active learning, but is probably sub-optimal as it concentrates solely on exploration rather than exploitation.
- **Uncertainty sampling:** Unlabelled examples closest to the current decision boundary are selected for labeling by the oracle. This strategy aims to rapidly acquire labels for those examples that are classified with least confidence. Note that maximum margin classifiers and boosting algorithms also aim to concentrate on patterns close to

the decision boundary, so it is perhaps not unreasonable to expect this strategy to perform well.

This gives three basic baselines, one with no active learning, one with a naïve active learning strategy, and one with a good active learning strategy.

3. Results

In this section, we present the results of experiments performed during the development phase of the challenge before moving on to describe the baseline submissions made on the final benchmark datasets.

3.1. Preliminary Experiments during the Development Phase

During the development phase of the challenge, a number of computationally intensive Monte-Carlo simulations were used to investigate the effectiveness of the three baseline active learning strategies. All of the labels made available for the training samples from each of the development datasets were downloaded. This allowed re-sampling to be used to estimate the variability in the performance of different active learning strategies due to the sample of data and due to any stochastic component of the learning procedure. For all experiments 100 replications were performed, each using a random partition of the available data to form training and test sets in the proportion of 3:1, and a positive example chosen at random from the training set as the "seed" pattern. The area under the receiver operating characteristic (ROC) curve (AUC) was recorded at approximately equal intervals on a logarithmic scale. The area under the resulting graph of AUC as a function of the number of labelled examples (on a logarithmic axis) then provides the test statistic, known as the area under the learning curve (ALC). Table 1 shows the ALC statistic for optimally regularised [kernel] ridge regression with passive, random sampling and uncertainty sampling active learning strategies. It can be seen that no active learning strategy is dominant, but more interestingly, random sampling is competitive with uncertainty sampling, even though it is a very naïve strategy.

The Friedman test, as recommended by Demšar (2006), reveals there is no significant difference in the average ranks of the three active learning strategies over the six development datasets. The lack of a significant difference is illustrated by the critical difference diagram, shown in Figure 1, which shows the average ranks of the three strategies, with the bar linking together cliques of statistically similar classifiers.

Figure 2 shows the average learning curves for the three baseline active learning strategies over the development benchmark datasets. Clearly active rather than passive learning is more useful on some datasets (NOVA, IBN_SINA and SYLVA) than others, such as HIVA, ORANGE and ZEBRA, where relatively little can be usefully learned from a small number of training patterns, whether they are selected at random or according to uncertainty.

3.2. Why does Random Active Learning Work so Well?

Figure 3 shows quantiles of the distribution of learning curves for the `nova` and `zebra` benchmarks, for random and uncertainty sampling active learning methods. It can be seen

Table 1: Area under the learning curve for three simple active learning strategies for the development datasets. The results are given as the arithmetic mean, and their standard errors, calculated over 100 random replications of the experiment. The best results for each dataset are shown underlined, without implication of statistical significance.

Benchmark	Passive	Random	Uncertainty
HIVA	0.2997 ± 0.0018	0.2505 ± 0.0056	0.1536 ± 0.0077
NOVA	0.4899 ± 0.0001	0.6975 ± 0.0033	0.6999 ± 0.0064
IBN_SINA	0.4821 ± 0.0002	0.8017 ± 0.0045	0.7832 ± 0.0050
ORANGE	0.2920 ± 0.0017	0.1910 ± 0.0052	0.2227 ± 0.0057
SYLVA	0.4967 ± 0.0000	0.8612 ± 0.0037	0.8893 ± 0.0025
ZEBRA	0.2744 ± 0.0013	0.3564 ± 0.0095	0.2949 ± 0.0120

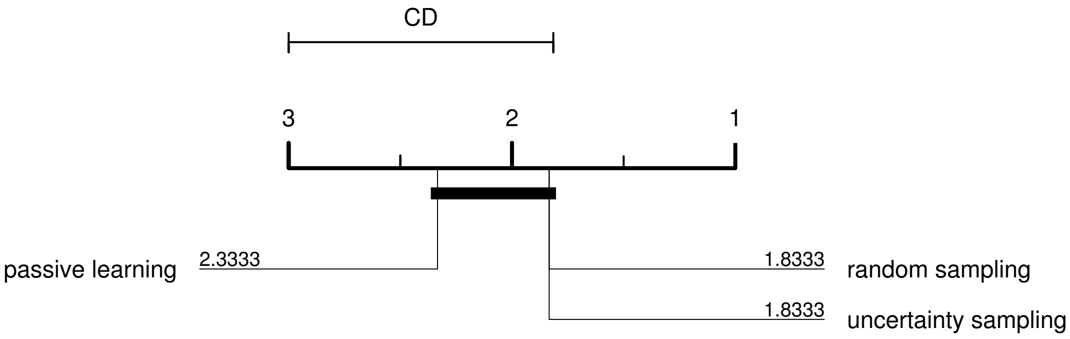


Figure 1: Critical difference diagram, showing the mean ranks of three basic active learning strategies over the final test benchmark datasets. The bar labelled "CD" shows the difference in mean rankings required for a statistically significant difference in performance to be detected.

that the uncertainty sampling strategy out-performs random active learning for the nova dataset with more than about 20 labelled examples (c.f. Figure 2b), while for smaller labelled datasets, however, uncertainty sampling performs poorly. The lower quantiles ($p_{.05}$ and $p_{.25}$) shown in Figure 3 suggest this is because of a large variability in the early part of the learning curves for the uncertainty sampling strategy. We conjecture that the downside of a principled strategy to active learning is that the selection of examples for labeling by the oracle depends on the current model, so if poor selections were made at an early stage, this adversely affects the quality of subsequent selections and hence learning proceeds slowly. This is less evident for random sampling, which gets locked into a poor hypothesis rather less frequently.

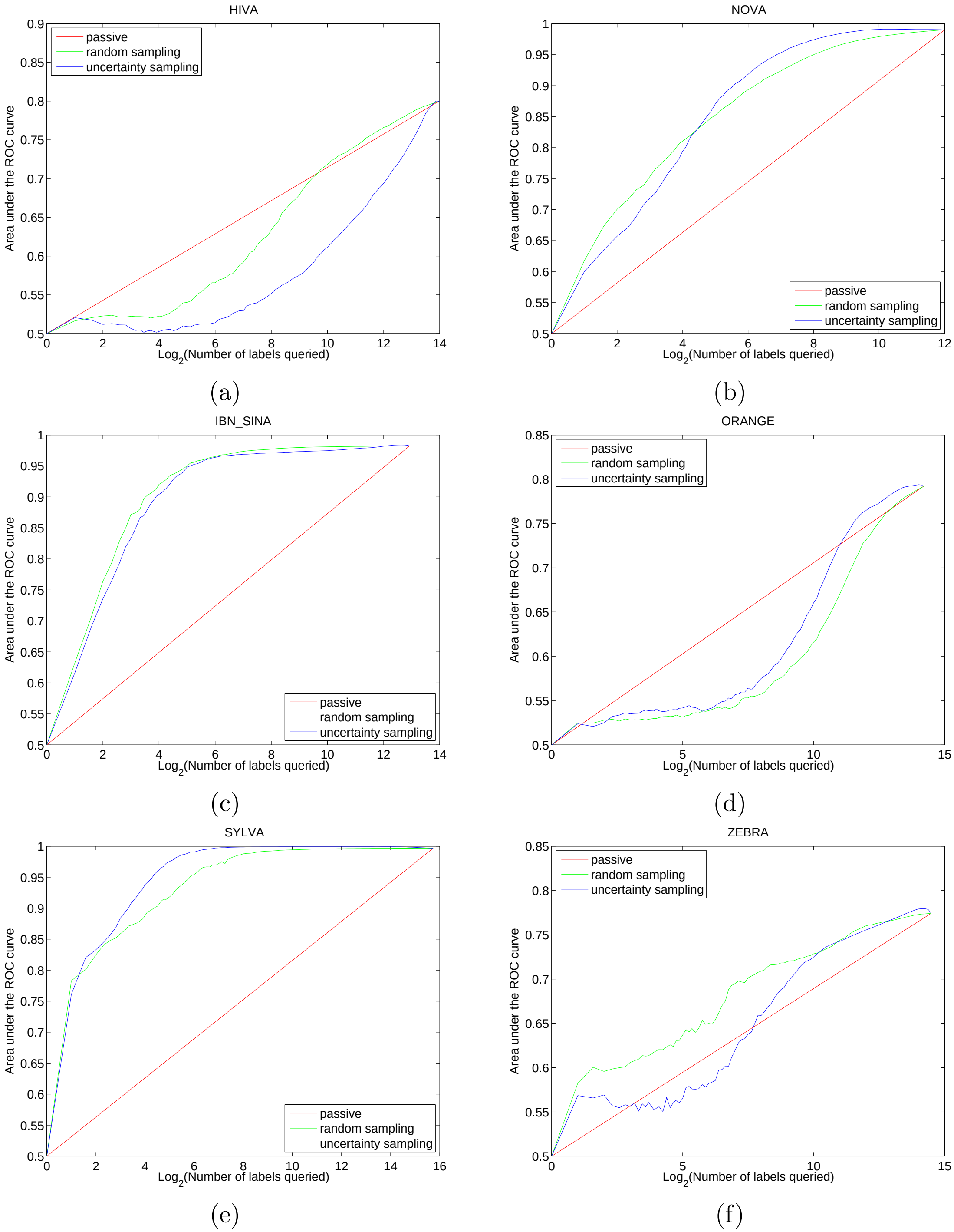
An effective active learning strategy must reach a near optimal trade-off between exploration and exploitation. The uncertainty sampling approach concentrates on exploiting the knowledge it has gained from the labels it has already acquired to further explore the decision boundary. The random sampling approach concentrates on exploration, and so is able to locate areas of the feature space where the classifier performs poorly. These results highlight the need for exploration as well as exploitation as the uncertainty sampling approach can become locked in a mistaken hypothesis of the location of the true decision boundary as it does not explore enough of the feature space that might suggest the current hypothesis is flawed.

3.3. Final Baseline Models

For the final test phase of the challenge, the baseline models were constructed according to the same protocol made available to the other participants (see Guyon et al., 2010, for details), and so Monte-Carlo simulations were not possible. A total of four baseline submissions were made using passive learning and random and uncertainty sampling based active learning. Two different initialization strategies were used: In the first, an initial classifier was constructed with the single positive seed pattern and the unlabelled patterns treated as if they belonged to the negative class. A second strategy was also used in conjunction with random sampling, where the prediction for unlabelled patterns was given by the Euclidean distance to the single positive pattern provided as a “seed” for the active learning procedure. This method would also have been used with the other active learning strategies had sufficient time been available, where the difference in initializations would have had a greater effect on the progress of the active learning procedure. The results obtained are shown in Table 2. The rankings of the baseline solutions show that a simple random sampling approach to active learning is effective and competitive with the results of some of the top submissions. The submission based on random sampling with linear initialization, for example, would have had an overall ranking of 4.667.

Again, the Friedman test was used to evaluate the statistical significance of any difference in the mean ranks of each approach, and again the differences were small, and not statistically significant. Figure 4 shows a critical difference diagram, illustrating the very similar rankings of the four baseline methods.

Figure 5 shows the learning curves obtained for the four baseline solutions for the six benchmark datasets used in the final phase of the challenge; the learning curve for the best submission for each benchmark is also shown. It can be seen that the results obtained for



(a)

(b)

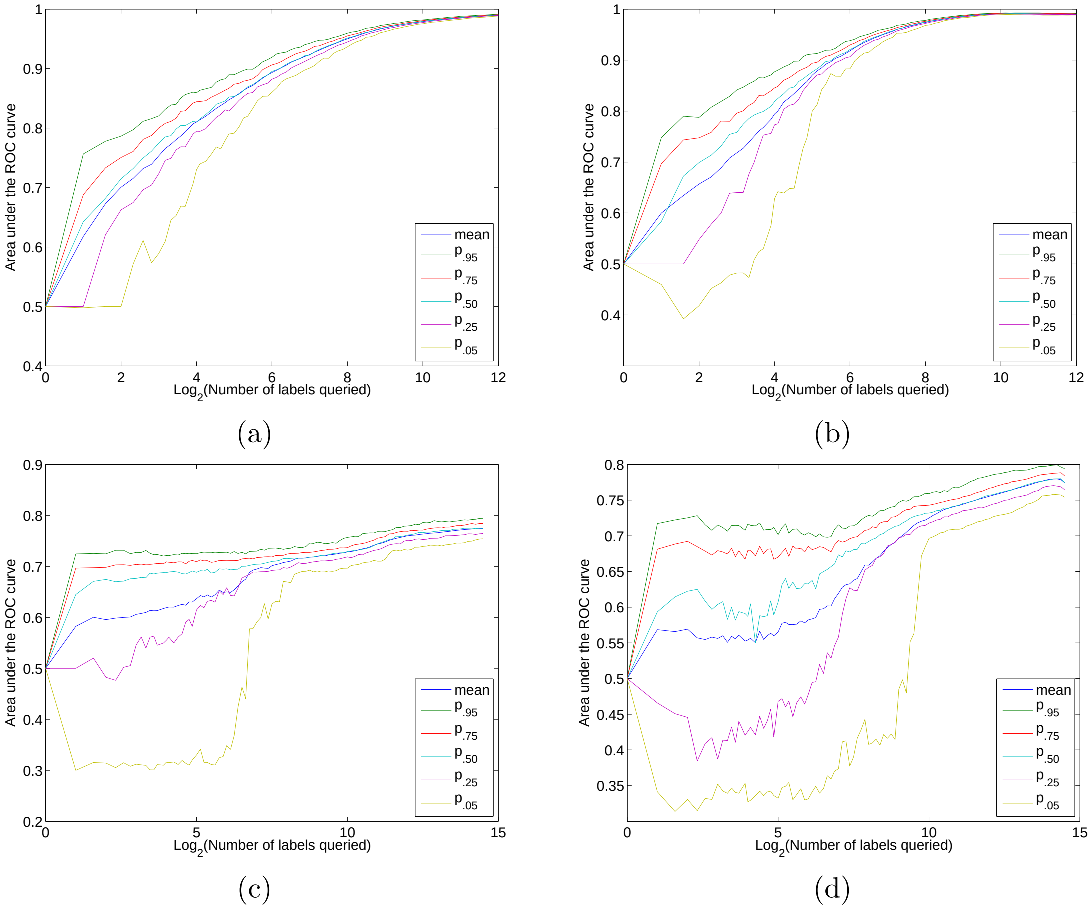
(c)

(d)

(e)

(f)

Figure 2: Average learning curves for active learning methods over 100 random realisations of the development benchmark datasets: (a) `hiva`, (b) `nova`, (c) `ibn_sina`, (d) `orange`, (e) `sylva` and (f) `zebra`.



(a)

(b)

(c)

(d)

Figure 3: Quantiles of the distribution of learning curves for random (a) and (c) and least certain (b) and (d) active learning methods over 100 random realisations of the `nova` (a) and (b) and `zebra` (c) and (d) development benchmark datasets.

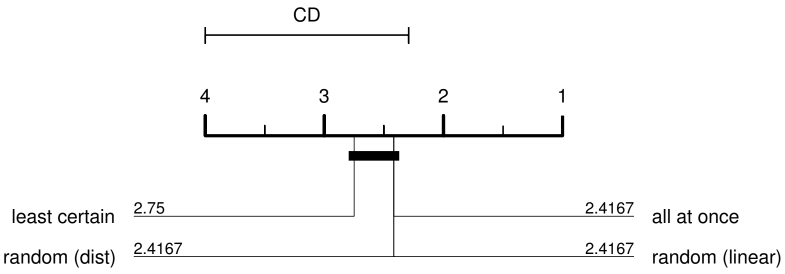


Figure 4: Critical difference diagram, showing the mean ranks of three basic active learning strategies over the final test benchmark datasets.

Table 2: Area under the Learning Curve (ALC) for the four baseline models and for the best entry for each of the final benchmark datasets. The best entries were as follows: A - `gcc4` (`reference`); B - `b` (`scan33scan33`); C - `C` (`chrisg`); D - `Dexp` (`datam1n`); E - `En` (`yukun`); F - `gccf2` (`reference`).

Method	**Global Score - ALC (rank)**					
	A	**B**	**C**	**D**	**E**	**F**
Passive	0.5455 (4)	0.3708 (3)	0.2663 (10)	0.4875 (21)	0.4966 (5)	0.7929 (5)
Random (linear)	0.5451 (5)	0.3084 (8)	0.2853 (6)	0.6512 (6)	0.4496 (8)	0.8217 (1)
Uncertainty sampling	0.4116 (15)	0.2689 (11)	0.2448 (11)	0.5748 (16)	0.3690 (16)	0.8074 (2)
Random (Euclidean)	0.6353 (1)	0.3195 (6)	0.3018 (5)	0.5996 (13)	0.4027 (12)	0.8048 (3)
Best	0.6353 (1)	0.3757 (1)	0.4273 (1)	0.8610 (1)	0.6266 (1)	0.8217 (1)

small numbers of labelled patterns are highly variable for all active learning methods for all benchmark datasets.

4. Summary

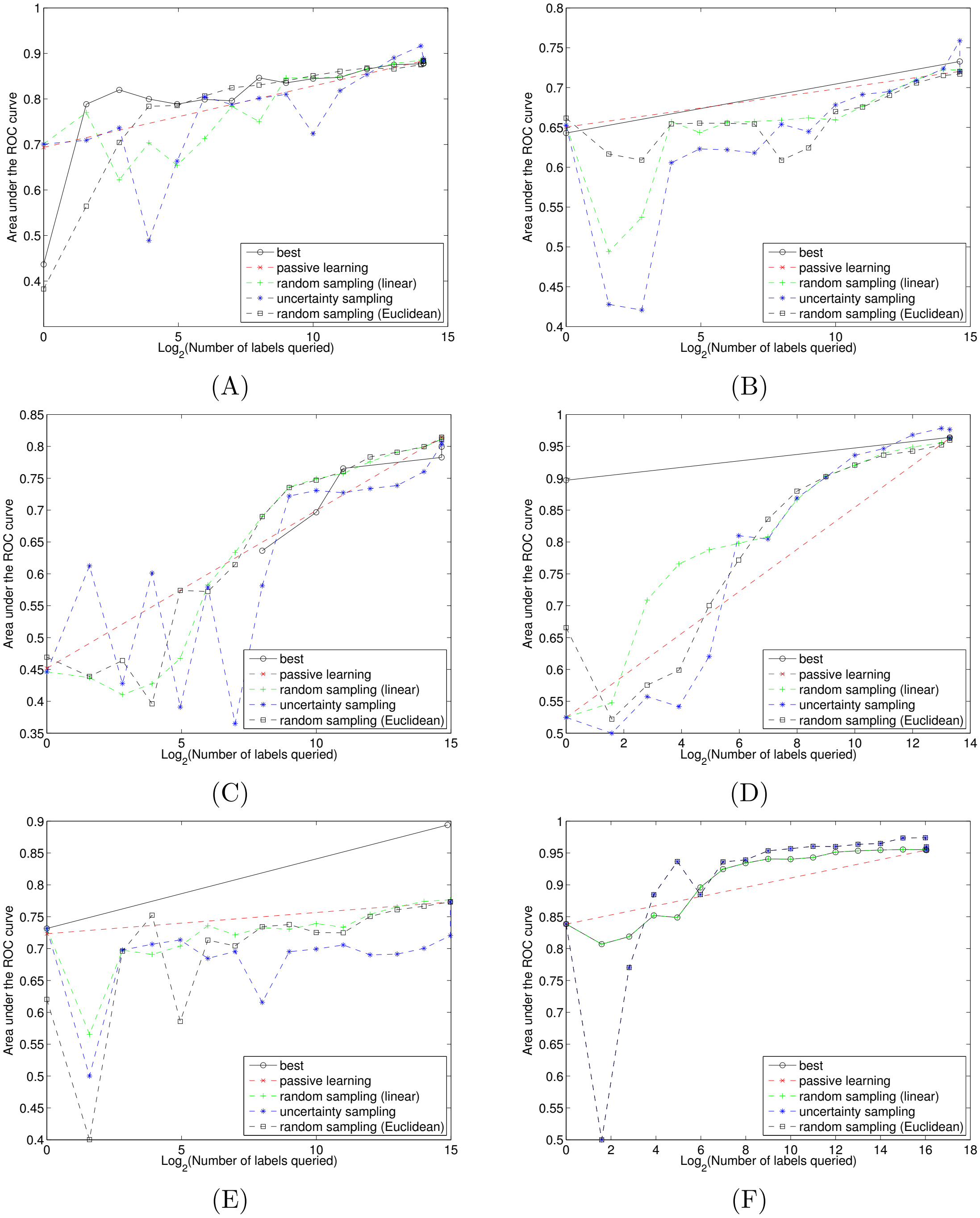
In this paper, we have described some simple baseline methods for the active learning challenge, based on optimally regularised ridge regression. A very basic random sampling approach was found to be competitive with both a more advanced uncertainty sampling approach and with some of the better challenge submissions. The poor performance of the uncertainty sampling approach seems likely to be due to a lack of exploration of the feature space at the expense of exploitation of current knowledge of the likely decision boundary. It is probable that better performance might be obtained using semi-supervised or transductive learning methods to take greater advantage of the availability of unlabelled data.

Acknowledgments

I would like to thank the anonymous reviewers for their helpful and constructive comments and the co-organizers of the challenge for their efforts in staging a very interesting and (for myself, at least ;o) educational challenge.

References

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



(A)

(B)

(C)

(D)

(E)

(F)

Figure 5: Learning curves for selected baseline models over the final benchmark datasets (A-F) of the active learning challenge.

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

G. C. Cawley. Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. In *Proceedings of the IEEE/INNS International Joint Conference on Neural Networks (IJCNN-06)*, pages 1661–1668, Vancouver, BC, Canada, July 16–21 2006. doi: 10.1109/IJCNN.2006.246634.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995. doi: 10.1007/BF00994018.

J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, January 1992. doi: 10.1162/neco.1992.4.1.1.

I. Guyon, G. Cawley, and G. Dror. Results of the active learning challenge. *Journal of Machine Learning Research, Workshop and Conference Proceedings*, 10, 2010.

J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, March 1986. doi: 0.1007/BF00116251.

K. Saadi, G. C. Cawley, and N. L. C. Talbot. Optimally regularised kernel Fisher discriminant classification. *Neural Networks*, 20(7):832–841, September 2007. doi: 10.1016/j.neunet.2007.05.005.

C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML-98)*, pages 515–521. Morgan Kaufmann, 1998.

B. Settles. Active learning literature survey. Technical Report 1648, School of Computer Sciences, University of Wisconsin-Maddison, 2009.

A. R. Webb. *Statistical pattern recognition.* Wiley, second edition, 2002.

S. Weisberg. *Applied linear regression.* Probability and Mathematical Statistics. John Wiley & Sons, second edition, 1985.