Design and Analysis of the WCCI 2010 Active Learning Challenge

Isabelle Guyon, Gavin Cawley, Gideon Dror, and Vincent Lemaire

Abstract— **We organized a data mining challenge on "active learning" for IJCNN/WCCI 2010, addressing machine learning problems where labeling data is expensive, but large amounts of unlabeled data are available at low cost. Examples include handwriting and speech recognition, document classification, vision tasks, drug design using recombinant molecules and protein engineering. Such problems might be tackled from different angles: learning from unlabeled data or active learning. In the former case, the algorithms must satisfy themselves with the limited amount of labeled data and capitalize on the unlabeled data with semi-supervised learning methods. Several challenges have addressed this problem in the past. In the latter case, the algorithms may place a limited number of queries to get new sample labels. The goal in that case is to optimize the queries and the problem is referred to as active learning. While the problem of active learning is of great importance, organizing a challenge in that area is non trivial. This is the problem we have addressed, and we describe our approach in this paper. The "active learning" challenge is part of the WCCI 2010 competition program (`http://www.wcci2010.org/competition-program`). The website of the challenge remains open for submission of new methods beyond the termination of the challenge as a resource for students and researchers (`http://clopinet.com/al`).**

I. Introduction

Much of the research on machine learning and data mining has so far concentrated on analyzing data that has already been collected, rather than on the collection of data. While experimental design is a well-developed discipline of statistics, data collection practitioners often neglect to apply such principled methods. As a result, data collected and made available to data analysts, in charge of explaining them and building predictive models, are not always of good quality and are often plagued by experimental artifacts. In reaction to this situation, some researchers in machine learning and data mining have started to become interested in experimental design to close the gap between data acquisition and experimentation and model building. This has given rise to the discipline of active learning.

From our perspective, to build good models, we need good data. However, collecting good data comes at a price. Experiments are usually expensive to perform and sometimes unethical or in more extreme cases impossible, while observational data are often available in abundance at a low cost. Practitioners must identify strategies for collecting data, which are both feasible and cost effective, resulting in the best possible models at the lowest possible cost. Hence, both efficiency and efficacy are important criteria in these evaluations. The setup of the active learning challenge considers sampling as the only intervention available to the data analyst or the learning machine, who may only place queries on the target values (labels).

In this challenge, we proposed several tasks involving pool-based active learning, where large unlabeled datasets are available from the outset of the challenge and the participants can place queries to acquire data for some amount of virtual cash. The participants were required to return prediction values for all the labels every time they purchased new labels. This allowed us to draw learning curves of prediction performance *vs.* the amount of virtual cash spent. The participants were judged according to the area under the learning curves, forcing them to optimize both efficacy (obtaining good predictive performance) and efficiency (spending little virtual cash).

The challenge consisted of 2 phases: a development phase (Dec. 1, 2009 - Jan. 31, 2010) during which the participants developed and tuned their algorithms using six development datasets and a final test phase (Feb. 3, 2010 - Mar. 10, 2010). Over 300 participants registered, downloaded the development data, experimented freely or tried making submissions on the website. The participants then regrouped into 30 teams, each of 1 to 20 members, which were manually verified. The teams competed for prizes on six new datasets provided for the final test phase. This level of participation is remarkable for a challenge that requires a deep level of commitment for participation because of specialized nature of the problem and the iterative submission protocol (participants must query for labels and make predictions by interacting with the website).

II. Brief overview of active learning

Modeling can have a number of objectives, including understanding or explaining the data, developing scientific theories, and making predictions. In this challenge we focus on predictive modeling, in a setup known in machine learning as "supervised learning". The goal is to predict an outcome y given a number of predictor variables $\mathbf{x} = [x_1, x_2, ...x_n]$, also called features, attributes, or factors. During training, the model (also called the learning machine) is provided with example pairs $\{\mathbf{x}, y\}$ (the training examples) with which to adjust its parameters. After training, the model is evaluated with new example pairs (the test examples) to estimate its generalization performance. In our framework, example pairs can only be obtained at a cost: optimal data acquisition must compromise between selecting many informative example pairs and incurring a large expense for data collection. Typically, either a fixed budget is available

Isabelle Guyon is an independent consultant. Direct correspondence to Clopinet, 955 Creston Road, Berkeley, CA 94708 (phone: +1 510 524 6211; email: isabelle@clopinet.com).

Gavin Cawley is with University of East Anglia, UK, Gideon Dror is with Academic College of Tel-Aviv-Yaffo, Israel, and Vincent Lemaire is with Orange, France.

and the generalization performance must be maximized or the data collection expenses must be minimized to reach or exceed a given generalization performance. Data pairs $\{\mathbf{x}, y\}$ are drawn identically and independently from an unknown distribution $P(\mathbf{x}, y)$. In the regular machine learning setting (passive learning), a batch of training pairs is made readily available from the outset. In the active learning setting, the labels y are withheld and can be purchased from an oracle. The learning machine must select the examples, which look most promising in improving the predictive power of the model. There exist several variants of active learning:

- **Pool-based active learning:** A large pool of examples of $\mathbf{x}$ is made available from the outset of training.
- **Stream-based active learning:** Examples are made available continuously.
- **De novo query synthesis:** The learner can select arbitrary values of x, i.e. use examples not drawn from $P(\mathbf{x})$.

Other scenarios, not considered here, include cases where data are not *i.i.d.*. Such situations occur in time series prediction, speech processing, unsegmented image analysis, and document analysis.

Of the variants of active learning considered, pool-based active learning is of considerable importance in current applications of machine learning and data mining, because of the availability of large amounts of unlabeled data in many domains, including pattern recognition (handwriting, speech, airborne or satellite images, etc.), text processing (internet documents, archives), chemo-informatics (untested molecules from combinatorial chemistry), and marketing (large customer databases). These are typical examples of the scenarios we want to study via the organization of this challenge. Stream-based active learning is also important when sensor data is continuously available and data cannot be easily stored. However, it is more difficult to evaluate in the context of a challenge, so we focus instead purely on pool-based active learning. Several of the techniques thus developed may also be applicable to stream-based active learning. The last type of active learning, "de-novo" query synthesis, will be addressed in upcoming experimental design challenges in which we will allow participants to intervene on $\mathbf{x}$. In this challenge, we limit the actions of the participants to sampling input space and query for y, we do not allow interventions on $\mathbf{x}$, such as setting certain values x_i.

A number of query strategies with various criteria of optimality have been devised. Perhaps the simplest and most commonly used query strategy is uncertainty sampling [1]. In this framework, an active learner queries the instances that it can label with least confidence. This of course requires the use of a model that is capable of assessing prediction uncertainty, such as a logistic regression model for binary classification problems. Another general active learning framework queries the labels of the instances that would impart the greatest change in the current model (expected model change), if we knew the labels. Since discriminative probabilistic models are usually trained with gradient-based optimization, the "change" imparted can be measured by the magnitude of the gradient [2]. A more theoretically motivated query strategy is query-by-committee (QBC) [3]. The QBC approach involves maintaining a committee of models, which are all trained on the current set of labeled samples, but represent competing hypotheses. Each committee member votes on the labels of query candidates and the query considered most informative is the one on which they disagree most. It can be shown that this is the query that potentially gives the largest reduction in the space of hypotheses (models) consistent with the current training dataset (version space). A related approach is Bayesian active learning. In the Bayesian setting, a prior over the space of hypotheses gets revised into a posterior after seeing the data. Bayesian active learning algorithms, for instance, Tong and Koller [4] maximize the expected Kullback-Leibler divergence between the revised posterior distribution (after learning with the new queried example) and the current posterior distribution given the data already seen. Hence this can be seen both as an extension of the expected model change framework for a Bayesian committee and a probabilistic reduction of hypothesis space. A more direct criterion of optimality seeks queries that are expected to produce the greatest reduction in generalization error, i.e. the error on data not used for training drawn from $P(\mathbf{x}, y)$ (expected error reduction). Cohn and collaborators [5] proposed the first statistical analysis of active learning, demonstrating how to synthesize queries that minimize the learner's future error by minimizing its variance. However, their approach applies only to regression tasks and synthesizes queries de novo. Another more direct, but computationally very expensive approach is to tentatively add to the training set all possible candidate queries with one of the opposite label and estimate how much generalization error reduction would result by adding it to the training set [6]. It has been suggested that uncertainty sampling and QBC strategies are prone to querying outliers and therefore are not robust. The information density framework [7] addresses that problem by considering instances that are not only uncertain, but representative of the input distribution, to be the most informative. This last type of approach addresses the problem of monitoring the trade-off between exploration and exploitation.

Several authors report that methods like "uncertainty sampling" often yield mediocre results because they stress only "exploitation". Conversely, "random sampling" relies only on "exploration". Methods compromising between exploration and exploitation usually perform best. These observations were confirmed in the challenge, as we will see in section V. For a more comprehensive survey, see [8].

III. Challenge Datasets

One of the exciting aspects of the organization of this challenge has been the abundance of data, which clearly signals that this problem is ripe for study, and solving it will have immediate impact. Several practitioners in need of good active learning methods offered to donate data from their

TABLE I

DEVELOPMENT DATASETS

Dataset	Domain	Feat. type	Feat. num.	Sparsity (%)	Missing (%)	Pos. lbls (%)	Tr & Te num.
ALEX	Toy problem	binary	11	0	0	72.98	5000
HIVA	Chemoinformatics	binary	1617	90.88	0	3.52	21339
IBN SINA	Handwriting rec.	mixed	92	80.67	0	37.84	10361
NOVA	Text ranking	binary	16969	99.67	0	28.45	9733
ORANGE	Marketing	mixed	230	9.57	65.46	1.78	25000
SYLVA	Ecology	mixed	216	77.88	0	6.15	72626
ZEBRA	Embryology	continuous	154	0.04	0.0038	4.58	30744

TABLE II

FINAL TEST DATASETS

Dataset	Domain	Feat. type	Feat. num.	Sparsity (%)	Missing (%)	Pos. lbls (%)	Tr & Te num.
Avicenna	Handwriting rec.	mixed	92	79.02	0	13.35	17535
Banana	Marketing	mixed	250	46.89	25.76	9.14	25000
Chemo	Chemoinformatics	mixed	851	8.6	0	8.1	25720
Docs	Text ranking	binary	12000	99.67	0	25.52	10000
Embryo	Embryology	continuous	154	0.04	0.0004	9.04	32252
Forest	Ecology	mixed	12	1.02	0	7.58	67628

study domain; we briefly describe these application domains and then summarize the data statistics.

A. Application Domains

We selected six different application domains, illustrative of the fields in which active learning is applicable:

Embryology: The problem is to study the development of living organisms, and in particular vertebrates. Modern fluorescence techniques allow the tracking of individual cell divisions and thus trace the development of whole organisms. One sub-problem is to identify cells that are initiating division (meiosis). This can be framed as a two-class classification problem (quiescent vs. dividing). Massive amounts of unlabeled data are available, but human expert labeling is very expensive and time consuming. Two model organisms (the zebrafish and the urchin) are under study as part of the FP6 Europen Community projects Embryomics (NEST adventure no. 12916) and BioEmergences (NEST no. 28809) at the Institute for Complex Systems in Paris (ISCPIF), France, http://iscpif.fr/tiki-index.php. Emmanuel Faure of ISCPIF assisted us in preparing data suitable for the challenge.

Chemoinformatics: Pharmacologists are on a constant quest for new small molecules, which are active against disease, yet are non-toxic. Screening large libraries of molecules in living organisms or even in-vitro is very costly. Hence, anything that can be done to prioritize experiments cuts down drug development costs. Chemoinformatics tackles the problem by applying machine learning techniques to molecular descriptors computed from a three dimensional description of the molecules and used to predict activity or toxicity. For example, a descriptor may be the number of carbon atoms, the presence of an aliphatic cycle, the length of the longest saturated chain, etc. A few examples of molecules having actually been tested in vitro are available for training and more real experimentation may be conducted (at additional expense). Large datasets are available for identifying compounds active against the HIV virus, tuberculosis or other diseases and for assessing the toxicity of kinases from PubChem http://pubchem.ncbi.nlm.nih.gov/. Curt Breneman, Professor in the Department of Chemistry and Chemical Biology, Director of Rensselaer Exploratory Center for Cheminformatics Research, at the Rensselaer Polytechnic Institute, Troy, New York, and his student Charles Bergeron generated molecular descriptors for the datasets we have identified together as most suitable. The data pre-processing was designed in collaboration with Kristin Bennett, professor in the Department of Mathematical Science and Department of Computer Sciences, at the same institute.

Handwriting recognition: Historical archive collections are difficult to process by traditional Optical Character Recognition (OCR) methods, due to their historical character types or due to the fact that the material is handwritten and use scripts that are no longer in use. There are thousands of different scripts in use worldwide and large volumes of scanned documents waiting to be indexed to facilitate retrieval. Active learning methods would accelerate the improvement of handwriting recognizers by making better use of the time of human experts to label data. Professor Mohamed Cheriet, Director of Synchromedia, Laboratory for multimedia communication in telepresence, École de

Technologie Supérieure, University of Quebec, Montreal, Quebec, Canada, and his students have prepared a large corpus of historical arabic documents for this challenge.

Text ranking: Internet search engines process billions of queries daily in order to rank web pages. The ever increasing number of documents available on the Internet makes this task ever more difficult. Ranking based on the relevance of content alone was replaced by a combination of contents and popularity (using information derived from web links). Now several dozens of features are commonly used by search engines. The problem remains of using these features in an optimal way for predicting the most satisfactory ranking, given a particular query. Very few labeled data are available for this task, but millions of documents are available. We will used Internet documents to illustrate these tasks.

Ecology: The state of the world is constantly monitored from space via satellite images. Airborne imaging systems also allow monitoring vegetation and activity. These massive amounts of data need to be processed automatically to assist experts in ecology, geography, geology, climatology, archaeology, and seismology. It is obviously costly to fully label such data by hand. We used data from the US Forest service, illustrative of such tasks.

Marketing: Consumers are tired of being constantly bombarded with advertisements, most of which are irrelevant to their needs. Targeted marketing aims at selectively directing marketing material to customers who are most likely to respond, by identifying their needs and susceptibility to advertising. Very few labeled data are available because it is costly to experiment with marketing strategies and to track consumer response. But large databases of customers are available. Vincent Lemaire of Orange, the French Telecom company, has kindly donated (with Orange's permission) a large marketing database, suitable for a challenge on active learning.

B. Statistics of the datasets

In most past challenges we organized [9–12], we used the same datasets during the development period and during the test period. Feed-back on a small "validation set" was provided during the development period and a larger final "test set" was reserved for the final evaluation. In the KDD cup 2009 [13], we explored the idea of having a separate development "toy" problem and perform the final testing on a different problem. We find the latter approach more satisfactory and it lends itself better to this particular challenge. Hence, we will use two sets of datasets, one for development and one for the final test. Although they may differ in feature representation, size and possibly in difficulty, they are drawn from similar domains to provide consistency: Embryology, chemoinformatics, handwriting recognition, text ranking, ecology, and marketing.

The statistics for the development datasets are shown in Table I and those of the final test datasets in Table II. We added an artificially generated toy dataset ALEX (Active Learning EXample) to the development datasets, in order to facilitate testing the sample code provided. The application domains and statistics on the labels for the final test datasets were kept confidential during the test phase and were revealed only at the end of the challenge. During the challenge, the final datasets were known only under their initial letter.

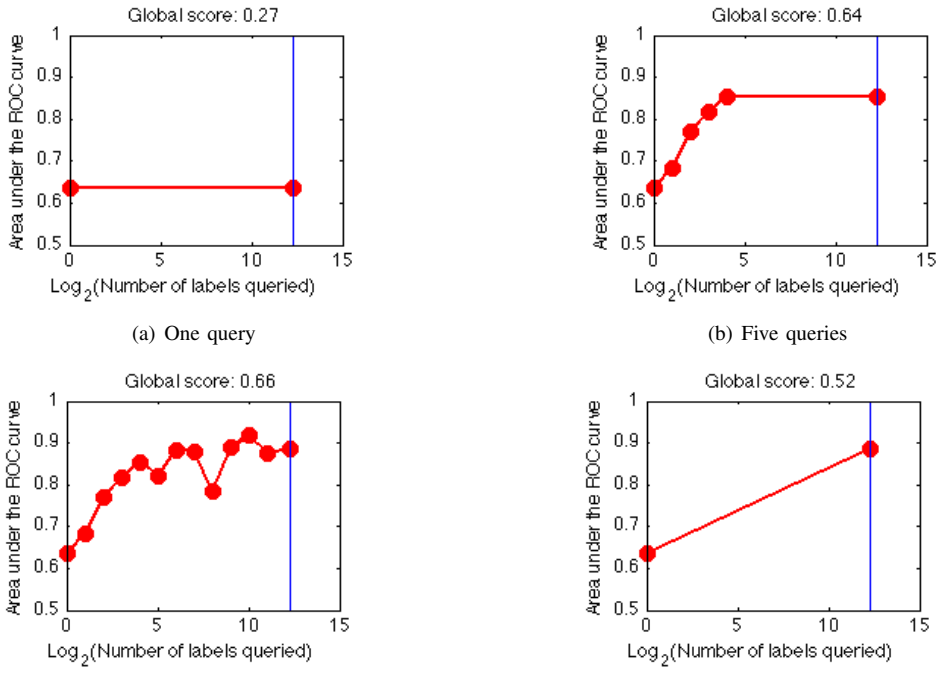
The problems chosen offered a wide range of difficulty levels, including heterogeneous noisy data (numerical and categorical variables), missing values, sparse feature representation, and unbalanced class distributions. To simplify the evaluation, we defined a two-class classification problem (binary labels) for each of these datasets. All datasets were split into training and test sets of identical sizes. We purposely chose only datasets with a large number of examples and a significant fraction of examples of the most depleted class, to obtain reasonable error bars on the results [14]. A detailed technical report on the datasets is also available [15].

IV. Protocol and evaluation

The protocol of the challenge was inspired by the previous competitions we have organized [16] and was designed to ensure fairness of the evaluation and stimulate participation. We give a brief synopsis of the rules and describe the evaluation metrics.

A. Rules

- **Goal of the challenge:** Given a data matrix of samples represented as feature vectors (samples in rows and features in columns), predict an unknown target variable (label). Initially a single example is labeled (the seed). The participants must predict all the labels as accurately as possible, by purchasing as few labels as possible.
- **Prizes:** To stimulate participation, we offered cash prizes and travel awards. Details are found on the website of the challenge.
- **Schedule:** The development period started on Dec. 1, 2009 and ended on Jan. 31, 2010. The final datasets were made available on Feb. 3, 2010. The challenge ended on Mar. 10, 2010 (with a one week extension from the original closing date).
- **Challenge protocol:** For each dataset, the participants were allotted a budget of "virtual cash" allowing them to "purchase" all the training data labels at the price of 1 ECU (experimental cash unit) per label. They made queries to the server by providing a list of samples for which they desire to purchase the label. Upon receipt of the labels, their account of virtual cash was debited. The participants were free to choose the number of queries and the number of samples per query. An experiment terminated when all the budget was spent or the challenge deadline was reached. To monitor progress, the participants were asked to provide predictions for all the labels every time they place a query, including the known and unknown labels of the training examples and the labels of the test examples.
- **Conditions of participation:** Anybody who complied with the rules of the challenge was welcome to participate. There was no requirement to disclose or publish methods.



(a) One query

(b) Five queries

(c) Thirteen queries

(d) Query all labels at once

Fig. 1. Example of learning curves for the toy dataset ALEX.

- **Anonymity:** The entrants were required to identify themselves to the organizers but could remain anonymous with respect to the outside world.
- **Team verification:** Towards the end of the development phase, the participants registered as teams. Each participant was allowed to enter only as part of a single team. The composition of the teams was manually verified. The team leaders were responsible for the team members conformance to the rules of the challenge.
- **Submission method:** The method of submission was by uploading results to the website, following the instructions provided.
- **Ranking:** During the development period, the scores were posted on a leaderboard on the website. The participants were allowed to perform multiple experiments on the same dataset, each time starting again with a new budget sufficient to purchase all the labels. During the final test period, no results were displayed until the challenge was over. Only one experiment per dataset per team was allowed. Separate rankings were performed for the various datasets.
- **Baseline results** The organizers uploaded baseline results to the website under the name"Reference". Those were not taken into account in determination of the winner of the challenge.
- **Reproducibility:** We forbade acquiring labels under fake names, by registering multiple times, or exchanging labels with other participants. Participation was not conditional on delivering code nor publishing methods. However, we asked the participants to voluntarily cooperate in reproducing their results. This included filling out a fact sheet about their methods and participating in post-challenge verification exercises.

For one of the datasets (dataset A), we provided a different set of target labels to each participant, without letting them know. In this way, if two teams exchanged labels, their poor performance should be suspicious. This would alert us, so that we could proceed with further checks, possibly asking the participants to provide their code. Our analysis of the performances on dataset A did not give us reason to suspect that anyone had cheated (see [17] for details). During the "verification phase" we asked the participants to redo their experiments on dataset A, this time providing the same labels to everyone. Those are the results provided in the result tables.

B. Evaluation metrics

1) The Area under the ROC Curve (AUC): The objective of the challenge is to make accurate predictions of the unknown values of the target variable (label), for the training examples not yet queried and the test examples. However, for the sake of simplicity, we ask participants to return prediction scores for all the samples, in the order of the patterns forming the data matrix. Prediction scores are not constrained to a specific range, but larger numerical values indicate higher confidence in positive class membership.

In many applications, tools producing scores are more usable than tools producing binary classifications. The participants were asked to provide a score (a discriminant value or a posterior probability $P(Y = 1|X)$), and were judged according to the area under the ROC curve (AUC).

The AUC is the area under the curve plotting sensitivity *vs.* (1− specificity) when the threshold θ is varied. We call "sensitivity" the error rate of the positive class and "specificity" the error rate of the negative class. The AUC is a standard metric in classification.

There are several ways of estimating error bars for the AUC. We used a simple heuristic, which gives us approximate error bars, and is fast and easy to implement: we find the point on the AUC curve corresponding to the largest balanced accuracy BAC = 0.5 (sensitivity + specificity). We then estimate the standard deviation of the BAC as:

$$\sigma = \frac{1}{2}\sqrt{\frac{p_+(1-p_+)}{m_+} + \frac{p_-(1-p_-)}{m_-}}\ , \qquad (1)$$

where m_+ is the number of examples of the positive class, m_- is the number of examples of the negative class, and p_+ and p_- are the probabilities of error on examples of the positive and negative class respectively, approximated by their empirical estimates, the sensitivity and the specificity [10].

2) Global Score: The Area under the Learning Curve: The prediction performance is evaluated according to the Area under the Learning Curve (ALC). A learning curve plots the Area Under the ROC curve (AUC) (see Section IV-B.1) computed on all the samples with unknown labels, as a function of the number of labels queried (including the seed). We consider two baseline learning curves:

- The ideal learning curve, obtained when perfect predictions are made (AUC=1). It goes up vertically then follows AUC=1 horizontally. It has the maximum area "$Amax$".
- The "lazy" learning curve, obtained by making random predictions (expected value of AUC: 0.5). It follows a straight horizontal line. We call its area "$Arand$".

To obtain our ranking score, we normalize the ALC as follows:

$$globalscore = (ALC - Arand)/(Amax - Arand)$$

We show in Figure 1(a), 1(b), and 1(c), learning curves for the toy example ALEX, obtained using the sample code, after 1, 5, and 13 queries. Each query acquired a different number of labels. We use a simple active learning strategy called "uncertainty sampling" (see Section II), with a very simple linear classifier making independence assumptions between variables.

We interpolate linearly between points and, for on-going experiments for which the entire budget has not yet been spent, we extrapolate the learning curve with a horizontal line. While participants could use all of their budget at once to purchase all the labels, a better global score (ALC) would probably be obtained by making incremental purchases. As an illustration, we show in Figure 1(d) the learning curve obtained with such a "passive" learning strategy consisting in purchasing all the labels at once; then a classifier is trained to produce prediction results using all the labeled examples.

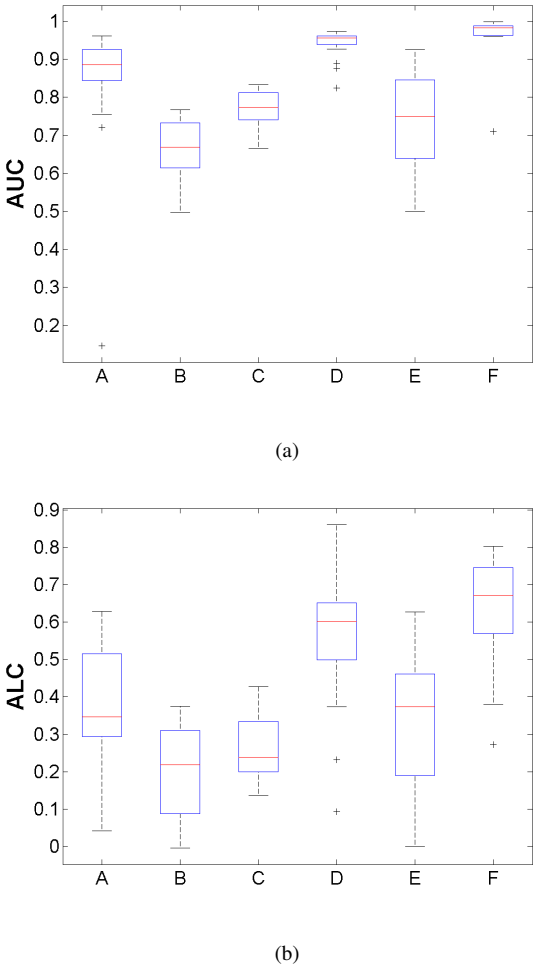


Fig. 2. **Distribution of results**. *We show box-whiskers plots for the various datasets. The red line represents the median, the blue boxes represent the quartiles, and the whiskers represent the range, excluding some outliers plotted individually as crosses. (a) Area under the ROC curve for the last point on the learning curve. (b) Area under the learning curve.*

The global score depends on how we scale the x-axis. We use a $\log_2$ scaling for all the datasets.

The participants are judged on the normalized ALC (global score). On-line feed-back on AUC and ALC performance was provided to the participants on development datasets only.

V. Challenge Results

It is difficult to make a fair assessment of the results on development data sets because the participants were allowed to perform multiple experiments on the same dataset; the knowledge of the labels obtained in previous experiments may have implicitly or explicitly be used in new experiments. Hence we report only results obtained on the final test sets for which the participants were only allowed to perform a single experiment. The distribution of performance with respect to AUC and ALC are shown in Figure 2. The results of the top ranking teams for each final dataset are found in Table III.

We encouraged the participants to enter results for multiple datasets by exponential scaling of the prizes with the number of wins. However, no team ended up winning on more than one dataset. The remaining prize money has been used to provide travel grants to encourage the winners to attend the workshop. For those participants who entered results on all 6 datasets, we performed a global ranking according to their average rank on the individual datasets. The overall winner by average rank (average rank 4.2) is the Intel team (Alexander Borisov and Eugene Tuv), who already ranked among the top entrants in several past challenges. The runner up by average rank (average ranked 4.8) is the ROFU team of National Taiwan University (Ming-Hen Tsai and Chia-Hua Hu). Other members of this research group headed by Chin en Lin have also won several previous machine learning challenges. The next best ranking teams are IDE (average rank 5.7) and Brainsignals (average rank 6.7). The team TEST (Zhili Wu) made entries on only 5 datasets, but did also very well (average rank 6.4).

The winning team (Intel) used a probabilistic version of the query-by-committee algorithm [18] with boosted Random Forest classifiers as committee members [19]. The batch size was exponentially increasing, disregarding the estimated model error. Some randomness in the selection of the samples was introduced by randomly sampling examples in a set of top candidates. No use was made of unlabeled data. This technique generated very smooth learning curves and reached high levels of accuracy for large numbers of training samples. The total run time on all development datasets on one machine is approximately 6-8 hours depending on model optimization settings. The method does not require any pre-processing, and naturally deals with categorical variables and missing values. The weakness of the method is at the beginning of the learning curve. Other methods making use of unlabeled data perform better in this domain.

The runner up team (ROFU) used Support Vector Machines (SVMs) [20] as a base classifier and a combination of uncertainty sampling and query-by-committee as the active learning strategy [21]. They made use of the unlabeled data [22, 23] and avoided sampling points near points already labeled. No active learning was performed on dataset B and E (inferring from the development dataset results that active learning would not be beneficial for such data). The method employed for learning from unlabeled data must not have been very effective because the results at the beginning of the learning curve are quite poor on some datasets (dataset C and F), but the performance for a large number of labeled examples are good. The authors report that using SVMs is fast so they could optimize the hyper-parameters by cross-validation.

Several participants found that uncertainty sampling and query-by-committee, without introducing any randomness in the selection process, may perform worse than random sampling. Query by committee performs better than uncertainty sampling both in randomized and non-randomized settings. Techniques for pro-actively sampling in regions with low densities of labels were reported not to yield significant improvements. Ensemble methods combined with query-by-committee active learning strategies yielded smooth learning curves. Good performances for very small number of examples ($\leq$ 100) were achieved only by teams using semi-supervised learning strategies.

A more detailed analysis of the challenge results is provided in [17].

TABLE III

RESULT TABLES FOR THE TOP RANKING TEAMS.

Dataset A		
Team	AUC (Ebar)	ALC
Flyingsky	0.8622 (0.0049)	0.6289
IDE	0.9250 (0.0044)	0.6040
ROFU	0.9281 (0.0040)	0.5533
JUGGERNAUT	0.8977 (0.0036)	0.5410
Intel	0.9520 (0.0045)	0.5273

Dataset B		
Team	AUC (Ebar)	ALC
ROFU	0.7327 (0.0034)	0.3757
IDE	0.7670 (0.0038)	0.3754
Brainsignals	0.7367 (0.0043)	0.3481
TEST	0.6980 (0.0044)	0.3383
Intel	0.7544 (0.0044)	0.3173

Dataset C		
Team	AUC (Ebar)	ALC
Brainsignals	0.7994 (0.0053)	0.4273
Intel	0.8333 (0.0050)	0.3806
NDSU	0.8124 (0.0050)	0.3583
IDE	0.8137 (0.0051)	0.3341
MUL	0.7387 (0.0053)	0.2840

Dataset D		
Team	AUC (Ebar)	ALC
DATAM1N	0.9641 (0.0033)	0.8610
Brainsignals	0.9717 (0.0033)	0.7373
ROFU	0.9701 (0.0032)	0.6618
TEST	0.9623 (0.0033)	0.6576
TUCIS	0.9385 (0.0037)	0.6519

Dataset E		
Team	AUC (Ebar)	ALC
DSL	0.8939 (0.0039)	0.6266
ROFU	0.8573 (0.0043)	0.5838
IDE	0.8650 (0.0042)	0.5329
Brainsignals	0.9090 (0.0039)	0.5267
Intel	0.9253 (0.0037)	0.4731

Dataset F		
Team	AUC (Ebar)	ALC
Intel	0.9990 (0.0009)	0.8018
NDSU	0.9634 (0.0018)	0.7912
DSL	0.9976 (0.0009)	0.7853
IDE	0.9883 (0.0013)	0.7714
DIT AI	0.9627 (0.0017)	0.7216

VI. CONCLUSIONS

The accumulation of massive amounts of unlabeled data and the cost of labeling have triggered a resurgence of interest in active learning. However, prior to this challenger, the newly proposed methods had never been evaluated in a fair contest. The challenge we organized for WCCI 2010 stimulated research in the field and delivered a comparative study free of "inventor bias". The analysis of the results reveals a

number of findings, which will need to be further validated by systematic experiments, in particular: the effectiveness of query-by-committee compared to uncertainty sampling; the benefit of introducing some degree of randomization in sampling; the edge obtained by combining active learning with semi-supervised learning (making use of unlabeled data) for the regime of small number of labeled examples (≤ 100). This challenge made use of our newly developed "virtual lab", which allows us to design interactive challenges in which the participants may query data generating processes. Having validated this methodology, this challenge opens the door to designing more advanced experimental design challenges in future.

ACKNOWLEDGMENTS

This project is an activity of the Causality Workbench supported by the Pascal network of excellence funded by the European Commission and by the U.S. National Science Foundation under Grant N0. ECCS-0725746. 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. Additional support was provided to fund prizes and travel awards by Microsoft, Orange FTP, and Health Discovery Corporation. We are very grateful to all the members of the causality workbench team for their contributions and in particular to our co-founders Constantin Aliferis, Greg Cooper, André Elisseeff, Jean-Philippe Pellet, Peter Spirtes, and Alexander Statnikov, and to the advisers and beta-testers Olivier Chapelle, Amir Reza Saffari Azar and Alexander Statnikov. The website was implemented by MisterP.net who provided exceptional support. This project would not have been possible without generous donations of data. We are very grateful to the data donors: Chemoinformatics – Charles Bergeron, Kristin Bennett and Curt Breneman (Rensselaer Polytechnic Institute, New York) contributed the dataset, used for final testing. Embryology – Emmanuel Faure, Thierry Savy, Louise Duloquin, Miguel Luengo Oroz, Benoit Lombardot, Camilo Melani, Paul Bourgine, and Nadine Peyriéras (Institut des systmes complexes, France) contributed the ZEBRA dataset. Handwriting recognition – Reza Farrahi Moghaddam, Mathias Adankon, Kostyantyn Filonenko, Robert Wisnovsky, and Mohamed Chériet (Ecole de technologie supérieure de Montréal, Quebec) contributed the IBN SINA dataset. Marketing – Vincent Lemaire, Marc Boullé, Fabrice Clérot, Raphael Féraud, Aurélie Le Cam, and Pascal Gouzien (Orange, France) contributed the ORANGE dataset, previously used in the KDD cup 2009. We also reused data made publicly available on the Internet. We are very grateful to the researchers who made these resources available: Chemoinformatics – The National Cancer Institute (USA) for the HIVA dataset. Ecology – Jock A. Blackard, Denis J. Dean, and Charles W. Anderson (US Forest Service, USA) for the SYLVA dataset (Forest cover type). Text processing – Tom Mitchell (USA) and Ron Bekkerman (Israel) for the NOVA datset (derived from the Twenty Newsgroups dataset).

REFERENCES

[1] D. Lewis and W. Gale, "A sequential algorithm for training text classifiers," in *ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM/Springer, 1994, pp. 3–12.

[2] B. Settles, M. Craven, and S. Ray, "Multiple-instance active learning," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 20. MIT Press, 2008, pp. 1289– 1296.

[3] H. S. Seung, M. Opper, and H. Sompolinskyy, "Query by committee," in *ACM Workshop on Computational Learning Theory*, 1992, pp. 287–294.

[4] S. Tong and D. Koller, "Active learning for parameter estimation in bayesian networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2000, pp. 647–653.

[5] D. Cohn, Z. Ghahramani, and M. Jordan, "Active learning with statistical models," *Journal of Artificial Intelligence Research*, vol. 4, pp. 129–145, 1996.

[6] N. Roy and A. McCallum, "Toward optimal active learning through sampling estimation of error reduction," in *International Conference on Machine Learning (ICML)*. Morgan Kaufmann, 2001, pp. 441–448.

[7] B. Settles, M. Craven, and S. Ray, "An analysis of active learning strategies for sequence labeling tasks," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, vol. 20. ACL Press, 2008, pp. 1069–1078.

[8] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.

[9] I. Guyon, S. Gunn, M. Nikravesh, and L. Z. Editors, *Feature Extraction, Foundations and Applications*, ser. Studies in Fuzziness and Soft Computing. With data, results and sample code for the NIPS 2003 feature selection challenge. Physica-Verlag, Springer, 2006.

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

[11] I. Guyon, A. Saffari, G. Dror, and G. Cawley, "Analysis of the IJCNN 2007 agnostic learning vs. prior knowledge challenge," in *Neural Networks*, vol. 21, Orlando, Florida, March 2008, pp. 544–550.

[12] I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov, "Design and analysis of the causation and prediction challenge," in *JMLR W&CP*, vol. 3, WCCI2008 workshop on causality, Hong Kong, June 3-4 2008, pp. 1–33. [Online]. Available: http://jmlr.csail.mit.edu/papers/topic/causality.html

[13] I. Guyon, V. Lemaire, M. Boullé, G. Dror, and V.Vogel, "Analysis of the kdd cup 2009: Fast scoring on a large orange customer database," in *JMLR W&CP*, vol. 7, KDD cup 2009, Paris, 2009.

[14] I. Guyon, J. Makhoul, R. Schwartz, and V. Vapnik, "What size test set gives good error rate estimates?" *PAMI*, vol. 20, no. 1, pp. 52–64, 1998.

[15] I. Guyon et al., "Datasets of the active learning challenge," Technical Report, 2010. [Online]. Available: http://clopinet.com/al/Datasets.pdf

[16] Clopinet, "Challenges in machine learning." [Online]. Available: http://clopinet.cm/challenges

[17] I. Guyon, G. Cawley, G. Dror, and V. Lemaire, "Results of the active learning challenge," in *JMLR W&CP*, Workshop on Active Learning and Experimental Design, collocated with AISTATS, Sardinia, Italy, May 16 2010.

[18] E. S. Y. Freund, H.S. Seung and N. Tishby, "Selective sampling using the query by committee algorithm," *Machine Learning*, vol. 28, pp. 133–168, 1997.

[19] A. Borisov, V. Eruhimov, and E. Tuv, "Tree-based ensembles with dynamic soft feature selection," in *Feature Extraction Foundations and Applications*, ser. Studies in Fuzziness and Soft Computing, I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, Eds., vol. 207. Springer, 2006.

[20] B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Fifth Annual Workshop on Computational Learning Theory*. Pittsburgh: ACM, 1992, pp. 144–152.

[21] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of Machine Learning Research*, 2002.

[22] S. K. V. Sindhwani, *Newton Methods for Fast Solution of Semi-supervised Linear SVMs*. MIT Press, 2005.

[23] V. Sindhwania and S. S. Keerthi, "Large scale semi-supervised linear svms," in *SIGIR*, 2006.