Datasets of the Active Learning Challenge

Isabelle Guyon ISABELLE@CLOPINETCOM

Report prepared by Isabelle Guyon with information from the data donors listed below:

Chemo-informatics (HIVA and C datasets): The National Cancer Institute (USA) provide the data used in the HIVA dataset. Charles Bergeron, Kristin Bennett and Curt Breneman (Rensselaer Polytechnic Institute, New York) contributed the C dataset.

Handwriting recognition (IBN-SINA and A datasets): Reza Farrahi Moghaddam, Mathias Adankon, Kostyantyn Filonenko, Robert Wisnovsky, and Mohamed Chériet (Ecole de technologie supérieure de Montréal, Quebec) contributed the datasets of Arabic manuscripts, IBN-SINA and A.

Text processing (NOVA and D datasets): - Tom Mitchell (USA) and Ron Bekkerman (Israel) provided the data used in the NOVA and D datasets (known as the Twenty Newsgroups).

Marketing (ORANGE and B datasets): Vincent Lemaire, Marc Boullé, Fabrice Clérot, Raphael Féraud, Aurélie Le Cam, and Pascal Gouzien (Orange, France) contributed the ORANGE and B datasets, previously used in the KDD cup 2009.

Ecology (SYLVA and F datasets): Jock A. Blackard, Denis J. Dean, and Charles W. Anderson (US Forest Service, USA) contributed the data used for the SYLVA and F datasets (Forest cover type).

Embryology (ZEBRA and E datasets): 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 and E datasets.

1. Introduction

Two times six datasets from various domains were made available for the Active Learning challenge (plus one toy dataset ALEX for practice purpose). The first six (Table 1) were made available during the development period. This gave the opportunity to the participants to practice without restriction and get performance feed-back on the results of their experiments from the on-line platform. Six other matching datasets (Table 2) were made available for final testing. Only one experiment could be made with the final datasets to enter the challenge.

Table 1: **Development datasets.** ALEX is a toy dataset given for illustrative purpose. The other datasets match the final datasets by application domain (see text).

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

Table 2: **Final test datasets.** The fraction of positive labels was not available to the participants.

Dataset	Domain	Feat. type	Feat. num.	Sparsity (%)	Missing (%)	Pos. lbls (%)	Tr & Te num.
A	Handwriting recog	mixed	92	79.02	0	13.35	17535
B	Marketing	mixed	250	46.89	25.76	9.14	25000
C	Chemo-informatics	mixed	851	8.6	0	8.1	25720
D	Text processing	binary	12000	99.67	0	25.52	10000
E	mbryology	continuous	154	0.04	0.0004	9.04	32252
F	Biology	mixed	12	1.02	0	7.58	67628

2. Data formats

All the data sets are in the same format and include 7 files in text format:

```
dataname.param    % Parameters and statistics about the data
dataname.data     % Unlabeled data (matrix of space delimited numbers,
                  % patterns in lines, features in columns).
dataname.mat      % The same data matrix in Matlab format.
dataname.label    % Target values.
dataname.labelid  % Identity of the labels (variables that are target
                  % values, i.e., columns of the label matrix.)
dataname.featid   % Identity of the features (variables that are not
                  % target values, i.e. columns of the data matrix)
dataname.dataid   % Identity of the samples (lines of the data matrix)
```

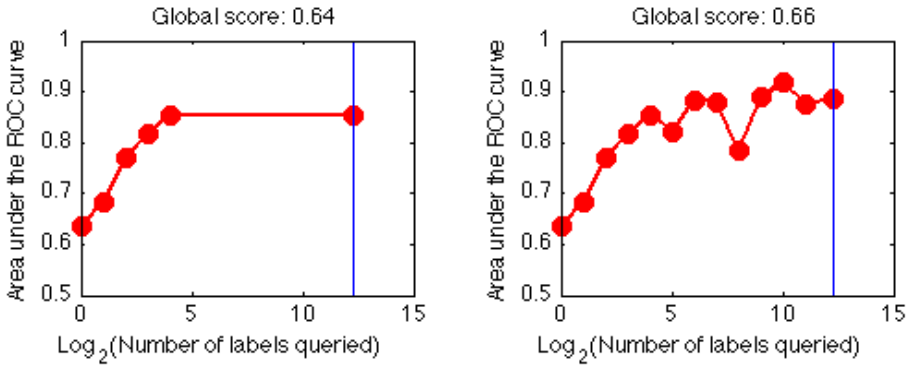
The participants used the following formats to send queries and results:

```
dataname.sample   % Sample numbers, one per line. Use to query labels.
dataname.predict  % Prediction values (as many as the total number of
                  % lines in the data matrix.
```

All problems were 2-class classification problems. The target classification values are therefore binary labels. All the unlabeled data (training and test data) were available from the

outset of the challenge. For each dataset, only one labeled training example was initially provided (called seed example). The rest of the labels for training examples were available for purchase for virtual cash from the challenge platform. The test labels were never disclosed.

The evaluation was performed by computing learning curves from predictions made by the participants: Every time a participant required a set of labels from the platform, he has to turn in predictions of class categories for all the examples. The Area under the ROC curve (AUC), a classical metric used to assess classification performance, was computed for examples not used yet for training (i.e. whose labels were not made available so far to the participant), including unlabeled training examples and test examples. The global scoring metric was the Area under the Learning Curve (ALC), appropriately normalized between 0 and 1: $global\_score = (ALC - Arand)/(Amax - Arand)$ where Arand is the expected value of the ALC for random predictions and Amax is the largest achievable ALC.



(a) (b)

Figure 1: Examples of learning curves for 5 and 13 experimental points.

A log scale in number of examples used for training was used. The learning curve was interpolated linearly between two experimental points and extrapolated horizontally to the total number of training examples whose labels were available for purchase (see Figure 2).

The datasets were incorporated in the Virtual Lab of the Causality Workbench `http://www.causality.inf.ethz.ch/workbench.php`. For each dataset, a wrapper was written in object-oriented Matlab to make is available in the GLOP package (Generative Lab Object Package) `http://www.causality.inf.ethz.ch//repository.php?id=23`

The names of the objects are: `@alex, @hiva, @ibn_sina, @nova, @orange, @sylva, @zebra, @A, @B, @C, @D, @E, @F`. Here is an example of using the GLOB package:

```
L=alex;                          % instantiate an alex model
data_profile(L);                 % show the dataset profile
label_profile(L);                % show the profile of the labels
task_n_pricing(L);               % show a description of the task
save_profile(L);                 % show the entire dataset profile
Qin=query(query_file);           % instantiate a query
[Qout, L]=process_query(L, Qin); % process query, return learning curve
```

In what follows, we present the design of the various datasets and show the learning curves produced either by the organizers using reference methods such as Least Square Support Vector Machines (LSSVM) and Selective Nave Bayes (SNB) or the overall winners (by average rank over all final evaluation datasets) Ideal Analytics, Intel, using gradient tree boosting. More details on these methods are found in JMLR W&CP volume 15. Note that these are not necessarily the best results. Other results can be viewed on the website of the challenge, which remains open for post-challenge submissions: `http://www.causality.inf.ethz.ch/activelearning.php?page=results#cont.`

3. Handwriting recognition: IBN_SINA and A datasets

3.1. Topic

The IBN_SINA and A datasets provides a feature representation of Arabic Historical Manuscripts. The letter A is mnemotechnical for Avicenna, the Latin name of the Arab scholar Ibn Sina.

3.2. Sources

- **Original owners:** The dataset is prepared on manuscript images provided by The Institute of Islamic Studies (IIS), McGill.

  Manuscript author: Abu al-Hasan Ali ibn Abi Ali ibn Muhammad al-Amidi (d. 1243 or 1233).

  Manuscript title: Kitab Kashf al-tamwihat fi sharh al-Tanbihat (Commentary on Ibn Sina's al-Isharat wa-al-tanbihat).

  Brief description: Among the works of Avicenna, his al-Isharat wa-al-tanbihat received the attention of the later scholars more than others. The reception of this work is particularly intensive and widespread in the period between the late twelfth century to the first half of the fourteenth century, when more than a dozen comprehensive commentaries on this work were composed. These commentaries were one of the main ways of approaching, understanding and developing Avicenna's philosophy and therefore any study of Post-Avicennian philosophy needs to pay specific attention to this commentary tradition. Kashf al-tamwihat fi sharh al-Tanbihat by Abu al-Hasan Ali ibn Abi Ali ibn Muhammad al-Amidi (d. 1243 or 1233), one of the early commentaries written on al-Isharat wa-al-tanbihat, is an unpublished commentary which still await scholars' attention.

- **Donors of the database:** Reza Farrahi Moghaddam, Mathias Adankon, Kostyantyn Filonenko, Robert Wisnovsky, and Mohamed Cheriet.

- **Contact:** Mohamed Cheriet - Synchromedia Laboratory
  ETS, Montréal, (QC) Canada H3C 1K3
  mohamed.cheriet@etsmtl.ca
  Tel: +1(514)396-8972
  Fax: +1(514)396-8595

- **Date received:** November 2009.

3.3. Reference

Reza Farrahi Moghaddam, Mohamed Cheriet, Mathias M. Adankon, Kostyantyn Filonenko, and Robert Wisnovsky. 2010. IBN SINA: a database for research on processing and understanding of Arabic manuscripts images. In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems (DAS '10). ACM, New York, NY, USA, 11-18.

3.4. Experimental design

The features were extracted following the procedure described in the JMLR W&CP paper: IBN SINA: A database for handwritten Arabic manuscripts understanding research, by Reza Farrahi Moghaddam, Mathias Adankon, Kostyantyn Filonenko, Robert Wisnovsky, and Mohamed Chériet. The data include 92 numeric features and 15 classes with at least 1000 positive examples. We created a number of binary classification problems for the development and final test datasets:

For IBN_SINA (development dataset), we selected the separation of the class aL vs. the rest.

For A (Avicenna, final evaluation dataset), we created 14 classification problems separating 2 classes vs. the rest. The two classes always included EU:

EU+nL
EU+qL
EU+bL
EU+lL
EU+tL
EU+kL
EU+vL
EU+fL
EU+mL
EU+rL
EU+hL
EU+dL
EU+yL

This allowed us to provide a seed example belonging to the class EU that was an example of the positive class for all 14 problems. We assigned at random a different problem to each participant. In this way we created a trap to catch eventual cheater who would exchange labels. After the challenge was over, we asked the participants to submit results again, this time all of them on the same problem.

3.5. Data statistics

See Tables 1 and 2. The samples in dataset A are different from those in IBN_SINA.

Table 3: Targets

Index	Name	Access	Type	Min	Max	FracPos (%)
1	EU	observable	binary	-1	1	6.53
2	aL	observable	binary	-1	1	37.84
3	bL	observable	binary	-1	1	5.18
4	dL	observable	binary	-1	1	4.86
5	fL	observable	binary	-1	1	5.79
6	hL	observable	binary	-1	1	10.57
7	kL	observable	binary	-1	1	5.39
8	lL	observable	binary	-1	1	24.85
9	mL	observable	binary	-1	1	13.16
10	nL	observable	binary	-1	1	12.71
11	qL	observable	binary	-1	1	5.43
12	rL	observable	binary	-1	1	5.88
13	tL	observable	binary	-1	1	5.42
14	vL	observable	binary	-1	1	13.75
15	yL	observable	binary	-1	1	13.96

Table 4: Variables

Index	**Name**	**Access**	**Type**	**Min**	**Max**
0	Target	observable	binary	-1	1
1	Aspect ratio	observable	continuous	0.039409	6.8387
2	Horizontal frequency	observable	categorical	1	12
3	Vertical CM ratio	observable	continuous	-0.94089	9.4262
4	Singular points	observable	categorical	0	24
5	Height ratio	observable	continuous	0.25714	5.8
6	Hole feature	observable	binary	0	1
7	End points	observable	categorical	0	15
8	Dot feature	observable	binary	0	1
9	BP_hole_1	observable	binary	0	1
10	BP_EP_1	observable	binary	0	1
11	BP_BP_1	observable	binary	0	1
12	BP_hole_2	observable	binary	0	1
13	BP_EP_2	observable	binary	0	1
14	BP_BP_2	observable	binary	0	1
15	BP_hole_3	observable	binary	0	1
16	BP_EP_3	observable	binary	0	1
17	BP_BP_3	observable	binary	0	1
18	BP_hole_4	observable	binary	0	1
19	BP_EP_4	observable	binary	0	1
20	BP_BP_4	observable	binary	0	1
21	BP_hole_5	observable	binary	0	1
22	BP_EP_5	observable	binary	0	1
23	BP_BP_5	observable	binary	0	1

Continued on next page

Table 4 – continued from previous page

Index	Name	Access	Type	Min	Max
24	BP_hole_6	observable	binary	0	1
25	BP_EP_6	observable	binary	0	1
26	BP_BP_6	observable	binary	0	1
27	EP_BP_1	observable	binary	0	1
28	EP_EP_1	observable	binary	0	1
29	EP_VCM_1	observable	categorical	0	2
30	EP_BP_2	observable	binary	0	1
31	EP_EP_2	observable	binary	0	1
32	EP_VCM_2	observable	categorical	0	2
33	EP_BP_3	observable	binary	0	1
34	EP_EP_3	observable	binary	0	1
35	EP_VCM_3	observable	categorical	0	2
36	EP_BP_4	observable	binary	0	1
37	EP_EP_4	observable	binary	0	1
38	EP_VCM_4	observable	categorical	0	2
39	EP_BP_5	observable	binary	0	1
40	EP_EP_5	observable	binary	0	1
41	EP_VCM_5	observable	categorical	0	2
42	EP_BP_6	observable	binary	0	1
43	EP_EP_6	observable	binary	0	1
44	EP_VCM_6	observable	categorical	0	2
45	BP_dot_UP_1	observable	binary	0	1
46	BP_dot_DOWN_1	observable	binary	0	1
47	BP_dot_UP_2	observable	binary	0	1
48	BP_dot_DOWN_2	observable	binary	0	1
49	BP_dot_UP_3	observable	binary	0	1
50	BP_dot_DOWN_3	observable	binary	0	1
51	BP_dot_UP_4	observable	binary	0	1
52	BP_dot_DOWN_4	observable	binary	0	1
53	BP_dot_UP_5	observable	binary	0	1
54	BP_dot_DOWN_5	observable	binary	0	1
55	BP_dot_UP_6	observable	binary	0	1
56	BP_dot_DOWN_6	observable	binary	0	1
57	EP_dot_1	observable	binary	0	1
58	EP_dot_2	observable	binary	0	1
59	EP_dot_3	observable	binary	0	1
60	EP_dot_4	observable	binary	0	1
61	EP_dot_5	observable	binary	0	1
62	EP_dot_6	observable	binary	0	1
63	Dot_dot_1	observable	binary	0	1
64	Dot_dot_2	observable	binary	0	1
65	Dot_dot_3	observable	binary	0	1
66	Dot_dot_4	observable	binary	0	1
67	Dot_dot_5	observable	binary	0	1
68	Dot_dot_6	observable	binary	0	1
69	EP_S_Shape_1	observable	categorical	0	2
70	EP_clock_1	observable	categorical	0	3
71	EP_UP_BP_1	observable	binary	0	1
72	EP_DOWN_BP_1	observable	binary	0	1
73	EP_S_Shape_2	observable	categorical	0	2

Continued on next page

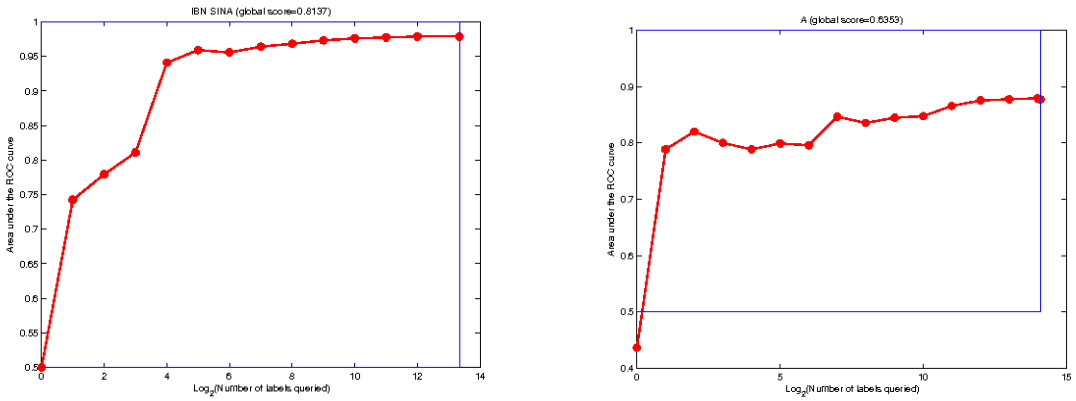
Table 4 – continued from previous page

Index	Name	Access	Type	Min	Max
74	EP_clock_2	observable	categorical	0	3
75	EP_UP_BP_2	observable	binary	0	1
76	EP_DOWN_BP_2	observable	binary	0	1
77	EP_S_Shape_3	observable	categorical	0	2
78	EP_clock_3	observable	categorical	0	3
79	EP_UP_BP_3	observable	binary	0	1
80	EP_DOWN_BP_3	observable	binary	0	1
81	EP_S_Shape_4	observable	categorical	0	2
82	EP_clock_4	observable	categorical	0	3
83	EP_UP_BP_4	observable	binary	0	1
84	EP_DOWN_BP_4	observable	binary	0	1
85	EP_S_Shape_5	observable	categorical	0	2
86	EP_clock_5	observable	categorical	0	3
87	EP_UP_BP_5	observable	binary	0	1
88	EP_DOWN_BP_5	observable	binary	0	1
89	EP_S_Shape_6	observable	categorical	0	2
90	EP_clock_6	observable	categorical	0	3
91	EP_UP_BP_6	observable	binary	0	1
92	EP_DOWN_BP_6	observable	binary	0	1

3.6. Baseline results

The balanced error rates (BER) for separating one class vs. all others were computed by training a Support Vector Machine (SVM) with kernel $K(x, y) = \exp(-\gamma d(x, y))$ $\gamma = 0.02$. Training and testing were done using the training and test sets of IBN_SINA.

Figure 2 depicts the results on IBN_SINA and dataset A with the reference method LSSVM produced by Gavin Cawley. See: `http://www.causality.inf.ethz.ch/activelearning.php?page=results#cont` for more results.



(a) (b)

Figure 2: Reference results by Gavin Cawley for IBN_SINA (a) and dataset A (b).

Table 5: Baseline results for IBN SINA

Problem no	Labels	SVM(BER)
1	EU	11.295
2	aL	3.671
3	bL	19.282
4	dL	10.657
5	fL	19.898
6	hL	7.057
7	kL	12.878
8	lL	7.66
9	mL	14.133
10	nL	16.663
11	qL	16.075
12	rL	14.875
13	tL	10.409
14	vL	8.57
15	yL	17.808

4. Marketing: ORANGE and B datasets

4.1. Topic

Customer Relationship Management (CRM) is a key element of modern marketing strategies. The datasets ORANGE and B (mnemotechnical for Banana) were extracted from a large marketing database from the French Telecom company Orange. The goal is to predict the propensity of customers to switch provider (churn), buy new products or services (appetency), or buy upgrades or add-ons proposed to them to make the sale more profitable (up-selling). The difficulties include heterogeneous noisy data (numerical and categorical variables), and unbalanced class distributions. For the ORANGE dataset (development dataset), we asked the participants to predict "appetency". For the B dataset (final evaluation dataset), we asked the participants to predict "[appetency OR upselling] AND NOT churn".

4.2. Source

The research team at Orange France who prepared the data includes Vincent Lemaire, Marc Boullé, Fabrice Clérot, Raphael Féraud, Aurélie Le Cam, and Pascal Gouzien. Contact: Vincent Lemaire vincent.lemaire@orange-ftgroup.com.

- **Donor of database:** This version of the database was prepared for the "Active Learning Challenge" by Isabelle Guyon, 955 Creston Road, Berkeley, CA 94708, USA (isabelle@clopinet.com).
- **Date received (original data):** November 2008, for the KDD cup 2009.

- **Date prepared for the challenge:** November 2009.

4.3. Past usage

The ORANGE dataset in a large and small version (more or less features) was used in the KDD cup 2009. Scoring was done using the Area under the ROC curve (AUC). The score is the average of the results on the 3 tasks (churn, appetency, and upselling). The best results (in score) were obtained by the IBM team, using the large dataset.

Table 6: Best performance for the tasks of Orange dataset

	Churn	Appetency	Upselling	Score
Fast track	0.7611	0.883	0.9038	0.849312
Slow track	0.7651	0.8819	0.9092	0.852062

For the small dataset, it is uncertain what the results are because some teams "unscrambled" the data and submitted large dataset results in lieu of small dataset results. The small dataset results were worse than the large dataset results. See for more details: Analysis of the KDD Cup 2009: Fast Scoring on a Large Orange Customer Database, Isabelle Guyon, Vincent Lemaire, Marc Boullé, Gideon Dror and David Vogel; JMLR W&CP 7: 1-22, 2009.

4.4. Experimental design

The following information was obtained from Orange: "A datamart of about one million Orange customers was used, with about ten tables and hundreds of fields. The first step was to resample the dataset, to obtain 100,000 instances with less unbalanced target distributions. For practical reasons (the challenge participants had to download the data), the same data sample was used for the three marketing tasks. In a second step, the feature construction language was used to generate 20,000 features and obtain a tabular representation. After discarding constant features and removing customer identifiers, we narrowed down the feature set to 15,000 variables (including 260 categorical variables). In a third step, for privacy reasons, data was anonymized, discarding variables names, randomizing the order of the variables, multiplying each continuous variable by a random factor and recoding categorical variable with randomly generated category name. To encourage participation, an easier task was also built from a reshuffled version of the datasets with only 230 variables." For the Active Learning challenge, we used the small dataset version with 230 variables. We randomly re-ordered the features and the examples. In addition, for dataset B, the features were disguised by random shifts and scaling for continuous values and by randomly assigning category values for categorical variables. Twenty "distracter" features were added using real variable whose values were randomly shuffled. These steps made it difficult to match the samples with the original data and guess the labels.

4.5. Number of examples and class distribution

The samples are the same in both datasets, but both samples and features are ordered differently. In addition the features in dataset B are disguised and some distracters have been added. See also Tables 1 and 2.

Fraction of positive examples (test and training sets):

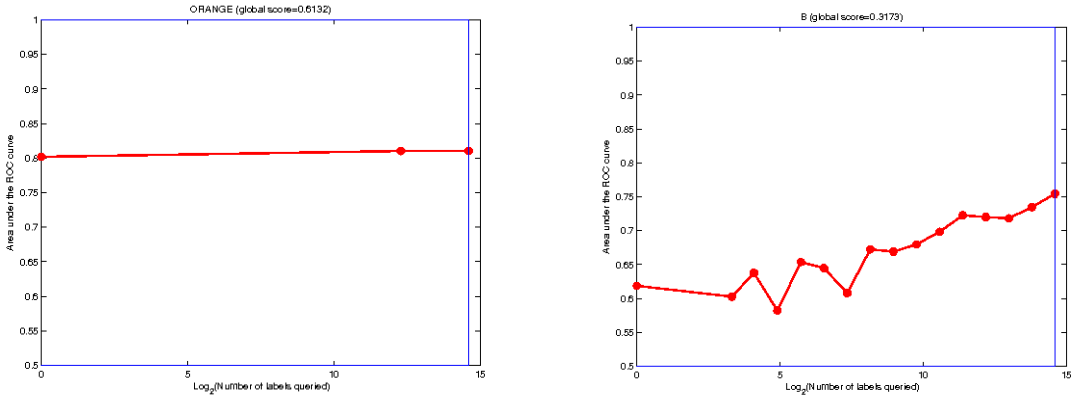
- Churn: 7.34
- Appetency: 1.78
- Upselling: 7.36

4.6. Type of input variables and variable statistics

Both continuous and categorical variables were found in data. There are 40 categorical variables and 190 continuous variables in the ORANGE data. Details can be obtained from the GLOP package by typing:
`save_profile(orange);`

4.7. Baseline results

The best reference results were produced by Marc Boullé using Selective Nave Bayes (SNB). We also show the results of the overall winners on dataset B. See: `http://www.causality.inf.ethz.ch/activelearning.php?page=results#cont` for more results.



(a) (b)

Figure 3: Reference results: March Boullé on the Orange dataset (a) and Ideal Analytics, Intel on the B dataset (b).

5. Ecology: SYLVA and F datasets

5.1. Topic

The tasks of the SYLVA and F datasets are to classify forest cover types. Both tasks were carved out of data from the US Forest Service (USFS). The task of SYLVA is to classify Ponderora pines vs. other classes of trees. The task of F is to classify Krummholz vs. other classes of trees.

5.2. Sources

- **Original owners:**
  Remote Sensing and GIS Program
  Department of Forest Sciences
  College of Natural Resources
  Colorado State University
  Fort Collins, CO 80523

  Jock A. Blackard
  USDA Forest Service
  3825 E. Mulberry
  Fort Collins, CO 80524 USA
  jblackard/wo_ftcol@fs.fed.us

  Dr. Denis J. Dean
  Associate Professor
  Department of Forest Sciences
  Colorado State University
  Fort Collins, CO 80523 USA
  denis@cnr.colostate.edu

  Dr. Charles W. Anderson
  Associate Professor
  Department of Computer Science
  Colorado State University
  Fort Collins, CO 80523 USA
  anderson@cs.colostate.edu

  Acknowledgements, Copyright Information, and Availability:
  Reuse of this database is unlimited with retention of copyright notice for Jock A. Blackard and Colorado State University.

- **Donor of database:**
  This version of the database was prepared for the "Active Learning Challenge" by Isabelle Guyon, 955 Creston Road, Berkeley, CA 94708, USA (isabelle@clopinet.com).

- **Date original data received:** August 28, 1998, UCI Machine Learning Repository, under the name Forest Cover Type.
- **Date prepared for the challenge:** November 2009.

5.3. Past usage

Blackard, Jock A. 1998. "Comparison of Neural Networks and Discriminant Analysis in Predicting Forest Cover Types." Ph.D. dissertation. Department of Forest Sciences. Colorado State University. Fort Collins, Colorado. Classification performance with first 11,340 records used for training data, next 3,780 records used for validation data, and last 565,892 records used for testing data subset: – 70% backpropagation – 58% Linear Discriminant Analysis. The subtask SYLVA was prepared for the WCCI 2006 "Performance Prediction Challenge" and the IJCNN 2007 "Agnostic Learning vs. Prior Knowledge" (ALvsPK) challenge is a 2-class classification problem. The best results were obtained with Logitboost by Roman Lutz with 0.4% balanced error rate (BER) in the PK track and 0.6% error in the AL track (`http://clopinet.com/isabelle/Projects/agnostic/Results.html`).

5.4. Experimental design

The original data comprises a total of 581012 instances (observations) grouped in 7 classes (forest cover types) and having 54 attributes (features) corresponding to 12 measures (10 quantitative variables, 4 binary wilderness areas and 40 binary soil type variables). The actual forest cover type for a given observation (30 x 30 meter cell) was determined from US Forest Service (USFS) Region 2 Resource Information System (RIS) data. Independent variables were derived from data originally obtained from US Geological Survey (USGS) and USFS data. Data is in raw form (not scaled) and contains binary (0 or 1) columns of data for qualitative independent variables (wilderness areas and soil types).

5.5. Variable Information

Given is the variable name, variable type, the measurement unit and a brief description. The forest cover type is the classification problem. The order of this listing corresponds to the order of numerals along the rows of the database.

Code Designations and Distribution

Wilderness Areas:

1. Rawah Wilderness Area
2. Neota Wilderness Area
3. Comanche Peak Wilderness Area
4. Cache la Poudre Wilderness Area

Soil Types: 1 to 40, based on the USFS Ecological Landtype Units for this study area.

Table 7: Variables of Sylva and F datasets

Name	Data Type	Units	Description
Elevation	quantitative	meters	Elevation in meters
Aspect	quantitative	azimuth	Aspect in degrees azimuth
Slope	quantitative	degrees	Slope in degrees
Horz_Distance_To_Hydr	quantitative	meters	Horz Dist to nearest surface water feature
Vert_Distance_To_Hydr	quantitative	meters	Vert Dist to nearest surface water feature
Horz_Distance_To_Rd	quantitative	meters	Horz Dist to nearest roadway
Hillshade_9am	quantitative	0 - 255	Hillshade index at 9am, summer solstice
Hillshade_Noon	quantitative	0 - 255	Hillshade index at noon, summer soltice
Hillshade_3pm	quantitative	0 - 255	Hillshade index at 3pm, summer solstice
Horz_Distance_To_FP	quantitative	meters	Horz Dist to nearest wildfire ignition point
Wilderness_Area (4 cols)	qualitative	0 /1	Wilderness area designation
Soil_Type (40 cols)	qualitative	0 /1	Soil Type designation
Cover_Type (7 types)	integer	1 to 7	Forest Cover Type designation

Table 8: Class codes and distribution

Name	code	number of records
Spruce/Fir	1	211840
Lodgepole Pine	2	283301
Ponderosa Pine	3	35754
Cottonwood/Willow	4	2747
Aspen	5	9493
Douglas-fire	6	17367
Krummholz	7	20501

Data preprocessing and data split

We carved a binary classification task out these data. For SYLVA Ponderosa pine is separated from all other trees and for F, Krummholz is separated from all other trees. For SYLVA, we created patterns containing the concatenation of 4 patterns: two of the target class and two randomly chosen from either class. In this way there are pairs of redundant features and of the features are non-informative. For F, we reverted to the original features, but recoded the categorical variables (Wilderness_Area and Soil_Type) with one variable taking integer values randomly assigned to the categories. We then randomized the order of the features and patterns and subsampled the patterns. In both cases, half of the data were reserved for training and half for testing.

5.6. Number of examples and class distribution

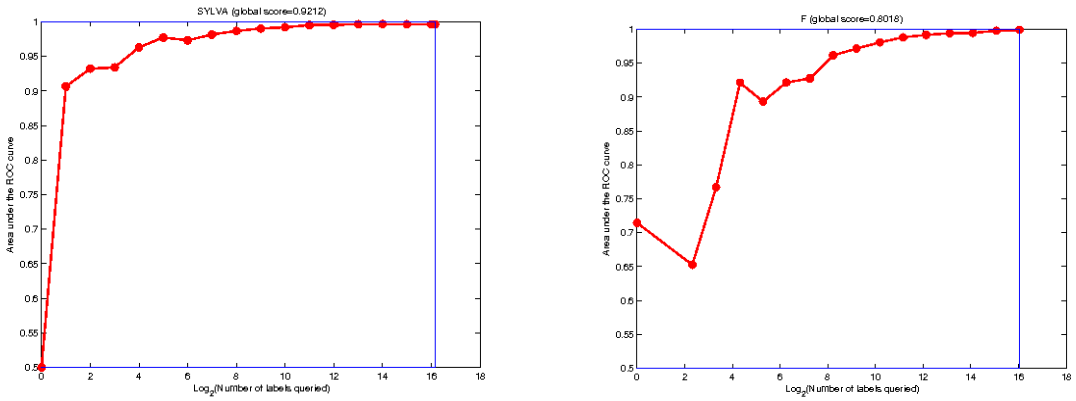
See Tables 1 and 2.

Table 9: Type of input variables and variable statistics

Index	Name	Access	Type	Min	Max
0	target	observable	binary	-1	1
1	Hillshade_Noon	observable	continuous	0	254
2	Soil_Type	observable	categorical	1	40
3	Slope	observable	continuous	0	65
4	Wilderness_Area	observable	categorical	1	4
5	Aspect	observable	continuous	0	360
6	Horizontal_Distance_To_Hydrology	observable	continuous	0	1397
7	Hillshade_9am	observable	continuous	0	254
8	Hillshade_3pm	observable	continuous	0	253
9	Vertical_Distance_To_Hydrology	observable	continuous	-166	599
10	Horizontal_Distance_To_Fire_Points	observable	continuous	0	7172
11	Horizontal_Distance_To_Roadways	observable	continuous	0	7117
12	Elevation	observable	continuous	1859	3858

5.7. Baseline results

We show below baseline results on SYLVA and the results obtained on the F dataset by the overall winners of the challenge. See: `http://www.causality.inf.ethz.ch/activelearning.php?page=results#cont` for more results.



(a) (b)

Figure 4: Reference results: Gavin Cawley on Sylva (a) and Ideal Analytics, Intel on the F dataset (b).

6. Chemo-informatics: HIVA and C datasets

6.1. Topic

The tasks of HIVA and C are to predict chemical activity of molecules. These are two-class classification problems. The variables represent properties of the molecule inferred from its structure. 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). For HIVA the task is to identify compounds that are active against the AIDS HIV infection. For the C dataset the problem is to predict the activation of pyruvate kynase, a well characterized enzyme, which regenerates ATP in glycolysis by catalyzing phosphoryl transfer from phosphoenol pyruvate to ADP to yield pyruvate and ATP. We next describe HIVA and C separately.

6.2. Sources

- **Original owners:**
  For the HIVA dataset, the data was made available by the National Cancer Institute (NCI), via the DTP AIDS Antiviral Screen program at: `http://dtp.nci.nih.gov/docs/aids/aids_data.html`. The DTP AIDS Antiviral Screen has checked tens of thousands of compounds for evidence of anti-HIV activity. Available are screening results and chemical structural data on compounds that are not covered by a confidentiality agreement.
- **Donor of database:**
  This version of the database was prepared for the "Active Learning Challenge" by Isabelle Guyon, 955 Creston Road, Berkeley, CA 94708, USA (isabelle@clopinet.com).
- **Date prepared for the challenge:** November 2009.

6.3. Past usage

An earlier release of the database was uses in an Equbits case study: `http://www.limsfinder.com/community/articles_comments.php?id=1553_0_2_0_C75`. The feature set was obtained by a different method. An earlier version of HIVA prepared by Isabelle Guyon for the WCCI 2006 "Performance Prediction Challenge" and the IJCNN 2007 "Agnostic Learning vs. Prior Knowledge" (ALvsPK) challenge. Depending on whether prior knowledge was used or not and depending on the dataset variants, the best performance was between 26% and 28% Balanced Error Rate (BER). The best result on HIVA in the WCCI 2006 Performance Prediction Challenge was obtained by Gavin Cawley (Test BER=0.275695, Test AUC=0.7671). See `http://clopinet.com/isabelle/Projects/agnostic/Results.html` for details.

6.4. Experimental design

The screening results of the May 2004 release containing the screening results for 43,850 compounds were used. The results of the screening tests are evaluated and placed in one of three categories:

- **CA - Confirmed active**
- **CM - Confirmed moderately active**
- **CI - Confirmed inactive**

We converted this into a two-class classification problem: Inactive (CI) vs. Active (CA or CM.) Chemical structural data for 42,390 compounds was obtained from the web page. It was converted to structural features by the program ChemTK version 4.1.1, Sage Informatics LLC. Four compounds failed parsing. The 1617 features selected include:

- unbranched_fragments: 750 features
- pharmacophores: 495 features
- branched_fragments: 219 features
- internal_fingerprints: 132 features
- ring_systems: 21 features

Only binary features having a total number of ones larger than 100 (¿400 for unbranched fragments) and at least 2% of ones in the positive class were retained. In all cases, the default program settings were used to generate keys (except for the pharmacophores for which "max number of pharmacophore points" was set to 4 instead of 3; the pharmacophore keys for Hacc, Hdon, ExtRing, ExtArom, ExtAliph were generated, as well as those for Hacc, Hdon, Neg, Pos.) The keys were then converted to attributes.

We briefly describe the attributes/features:

Branched fragments: each fragment is constructed through an "assembly" of shortest-path unbranched fragments, where each of the latter is required to be bounded by two atoms belonging to one or more pre-defined "terminal-atom".

Unbranched fragments: unique non-branching fragments contained in the set of input molecules.

Ring systems: A ring system is defined as any number of single or fused rings connected by an unbroken chain of atoms. The simplest example would be either a single ring (e.g., benzene) or a single fused system (e.g., naphthalene).

Pharmacophores: ChemTK uses a type of pharmacophore that measures distance via bond connectivity rather than a typical three-dimensional distance. For instance, to describe a hydrogen-bond acceptor and hydrogen-bond donor separated by five connecting bonds, the corresponding key string would be "HAcc.HDon.5". The pharmacophores were generated from the following features:

- Neg. Explicit negative charge.
- Pos. Explicit positive charge.
- HAcc. Hydrogen-bond acceptor.
- HDon. Hydrogen-bond donor.

- ExtRing. Ring atom having a neighbor atom external to the ring.
- ExtArom. Aromatic ring atom having a neighbor atom external to the ring.
- ExtAliph. Aliphatic ring atom having a neighbor atom external to the ring.

Internal fingerprints: small, fixed catalog of pre-defined queries roughly similar to the MACCS key set developed by MDL.

We matched the compounds in the structural description files and those in the compound activity file, using the NSC id number. We ended up with 42678 examples.

6.5. Number of examples and class distribution

See Table 1.

6.6. Type of input variables

All variables are binary. The data was saved as a non-sparse matrix, even though it is 91% sparse because dense matrices load faster in Matlab and the ASCII format compresses well.

6.7. Baseline results

We show below baseline results for HIVA using the reference method LSSVM. See: `http://www.causality.inf.ethz.ch/activelearning.php?page=results#cont` for more results.

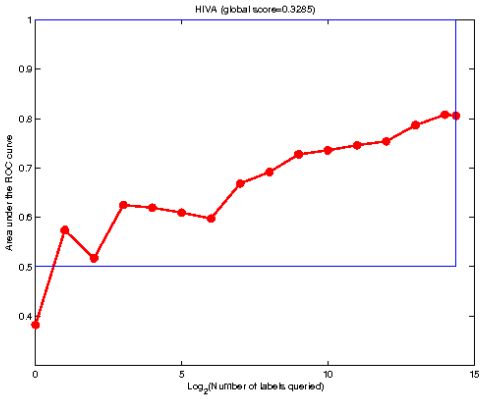


Figure 5: Reference results of Gavin Cawley on Hiva

7. The C (Chemo-informatics) dataset

7.1. Sources

- **Original owners:**
  The C dataset is a dataset for assessing the toxicity of kinases that was downloaded from PubMed `http://pubchem.ncbi.nlm.nih.gov/`. This dataset assayed pyruvate kinase in 51441 compounds, and the qHTS experimental results appear under assay identification number (AID) 361 on PubChem. Note that a Google search for '361' and '51441' returns a PubChem page as a top search result.

- **Donors of database:**
  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 students Micheal Krein and Charles Bergeron generated molecular descriptors for the datasets we have identified together as most suitable. The data preprocessing was designed in collaboration with Kristin Bennett, professor in the Department of Mathematical Science and Department of Computer Sciences, working at the same institute.

- **Date prepared for the challenge:** November 2009.

7.2. Past usage

None in the context of a challenge.

7.3. Experimental design

The data relates to drug discovery. The first step in drug design requires identifying a small number of screening hits that are effective at modulating a disease-specific biological pathway. Traditionally, a large number of compounds are assayed at a single concentration; this is called high-throughput screening (HTS), a mainstay of pharmaceutical development. A recent technique called quantitative high-throughput screening (qHTS) obtains more complete dose-response information by assaying compounds at multiple concentrations in a single experiment. The half-maximal activity concentration pAC_50 is the (negative log-10) concentration at which the midpoint of the activity range is attained.

This dataset assayed pyruvate kinase in 51441 compounds, and the qHTS experimental results appear under assay identification number (AID) 361 on PubChem. QSAR descriptors (MOE and TAE-RECON) were generated at RPI by Micheal Krein, a PhD student with Prof. Breneman. The substance identification (SID) that may be looked up on PubChem to obtain the molecular structures that are used to generate computational chemistry descriptors.

The team of prof. Breneman calculated their own pAC_50's that are more reliable than the ones reported on the PubChem website. Compounds displaying no activity over the tested concentration range are assigned pAC_50=0. The same is true for a small number of irregular samples that do not follow the expected dose-response behavior. These are

arbitrary choices, as 'some unknown number below $\simeq 3.5$' and 'some unknown real number' would be more accurate statements for inactive and irregular, respectively.

For a classification task, samples having pAC_50$\geq$ 4.94 are interpreted as screening hits and the others are not. PubChem suggests an intermediate category that I call 'junior screening hits' with 4.24 <pAC_50< 4.94. Most samples in this category have pAC_50's that are uncertain, a problem that is significantly improved by the new, more reliable method for calculating pAC_50's.

For the challenge, all positive values of pAC_50 were associated with a positive target value and the others with a negative target value.

7.4. Number of examples and class distribution

See Table 2.

7.5. Type of input variables

Most variables are continuous, some are binary. To obtain the full dataset profile from the GLOP package, type at the Matlab prompt:

```
save_profile(C);
```

7.6. Baseline results

We show the results on the C dataset from the overall challenge winners. See: `http://www.causality.inf.ethz.ch/activelearning.php?page=results#cont` for more results.

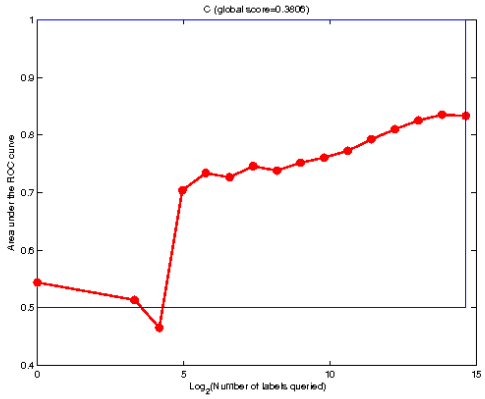


Figure 6: Results on the C dataset by Ideal Analytics, Intel.

8. Document classification: NOVA and the D dataset

8.1. Topic

The task of the NOVA and D datasets is Document classification from the 20-Newsgroup data. We selected the separation of politics and religion topics from all the other topics for NOVA and the separation of all newsgroups relating to computers vs. others for the D dataset. In both cases these are two-class classification problem with sparse binary input variables using a bag-of-word representation with a vocabulary of approximately 17000 words.

8.2. Sources

- **Original owners:**
  Tom Mitchell
  School of Computer Science
  Carnegie Mellon University
  tom.mitchell@cmu.edu
  Available from the UCI machine learning repository. The version we are using was preprocessed by Ron Bekkerman `http://www.cs.technion.ac.il/~ronb/thesis.html` into the "bag-of-words" representation.

- **Donor of database:**
  This version of the database was prepared for the Active Learning challenge on performance prediction by Isabelle Guyon, 955 Creston Road, Berkeley, CA 94708, USA (isabelle@clopinet.com).

- **Date prepared for the challenge:** November 2009.

8.3. Past usage

T. Mitchell. Machine Learning, McGraw Hill, 1997.
T. Joachims (1996). A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization, Computer Science Technical Report CMU-CS-96-118. Carnegie Mellon University.
Ron Bekkerman, Ran El-Yaniv, Naftali Tishby, and Yoad Winter. Distributional Word Clusters vs. Words for Text Categorization. JMLR 3(Mar):1183-1208, 2003.

Versions of NOVA were prepared for the WCCI 2006 "Performance Prediction Challenge" and the IJCNN 2007 "Agnostic Learning vs. Prior Knowledge" (ALvsPK). The best results were between 4% and 6% balanced error rate (BER), see `http://clopinet.com/isabelle/Projects/agnostic/Results.html`.

8.4. Experimental design

We selected 8 newsgroups relating to politics or religion topics as our positive class (Table C.1.) Vocabulary selection includes the following filters:

remove words containing digits and convert to lowercase

remove words appearing less than twice in the whole dataset.
remove short words with less than 3 letters.
exclude 2000 words found frequently in all documents.
truncate the words at a max of 7 letters.

Both NOVA and the D dataset come from the same original data. D is disguised compared to NOVA by removing a few features and adding distracters (random permutations of the original features). Some examples are repeated in D to make the size of the dataset different. Finally the order of the features and samples is randomized.

Table 10: Twenty Newsgroups categories, together with the positive/negative targets for the Nova and D datasets.

Newsgroup	Number of examples	Nova targets	D targets
alt.atheism	1114	+	-
comp.graphics	1002	-	+
comp.os.ms-windows.misc	1000	-	+
comp.sys.ibm.pc.hardware	1028	-	+
comp.sys.mac.hardware	1002	-	+
comp.windows.x	1000	-	+
misc.forsale	1005	-	-
rec.autos	1004	-	-
rec.motorcycles	1000	-	-
rec.sport.baseball	1000	-	-
rec.sport.hockey	1000	-	-
sci.crypt	1000	-	-
sci.electronics	1000	-	-
sci.med	1001	-	-
sci.space	1000	-	-
soc.religion.christian	997	+	-
talk.politics.guns	1008	+	-
talk.politics.mideast	1000	+	-
talk.politics.misc	1163	+	-
talk.religion.misc	1023	+	-

8.5. Number of examples and class distribution

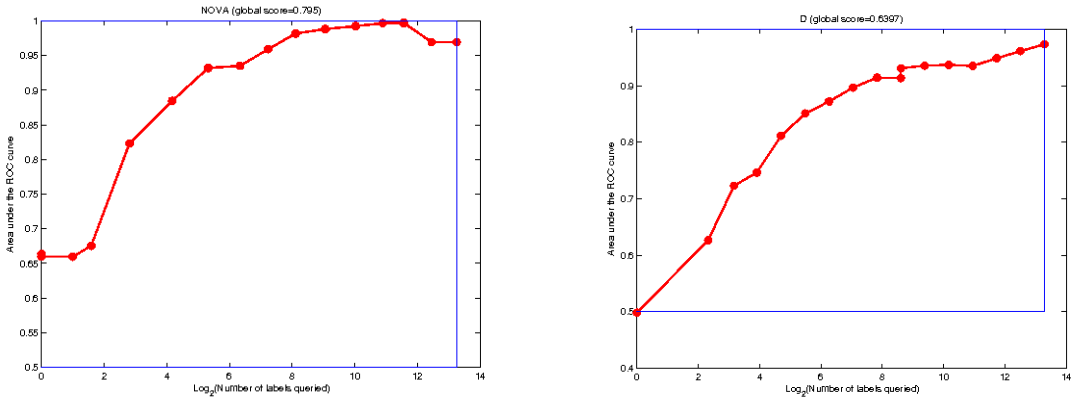
See Tables 1 and 2.

8.6. Type of input variables and variable statistics

All variables are binary. There are no missing values. The data is very sparse. Over 99% of the entries are zero. The data was saved as a sparse-binary matrix.

8.7. Baseline results

We show below sample results by the organizing team on NOVA and by the overall challenge winners on dataset D. See: `http://www.causality.inf.ethz.ch/activelearning.php?page=results#cont` for more results.



(a) (b)

Figure 7: Reference results on NOVA by Gideon Dror (a) and on dataset D by Ideal Analytics, Intel (b).

9. Embryology: ZEBRA and E datasets

9.1. Topic

The ZEBRA and E datasets provide a feature representation of cells of zebrafish embryo to determine whether they are in division (meiosis) or not. All the examples in this subset are manually annotated.

9.2. Sources

- **Original owners:**
  Emmanuel Faure, Thierry Savy, Louise Duloquin, Miguel Luengo Oroz, Benoit Lombardot, Camilo Melani, Paul Bourgine and Nadine Peyrieras.

  Acknowledgements, Copyright Information, and Availability
  Thanks to the Embryomics Consortium.
  Thanks to IN2P3 (French National Calcul Center)

- **Donor of database:**
  Emmanuel Faure.

- **Date received:**
  November 2009.

9.3. Past usage

Nothing with these features. Other versions of these datasets were used in various publications, including: Cell Lineage Reconstruction of Early Zebrafish Embryos Using Label-Free Nonlinear Microscopy, Nicolas Olivier, Miguel A. Luengo-Oroz, Louise Duloquin, Emmanuel Faure, Thierry Savy, Israël Veilleux, Xavier Solinas, Delphine Débarre, Paul Bourgine, Andrés Santos, Nadine Peyriéras and Emmanuel Beaurepaire Science 20 August 2010: 329 (5994), 967-971.

9.4. Experimental design (Emmanuel Faure)

Our Embryomics project is devoted to the morphodynamical "reconstruction" of the cell lineage tree underlying the processes of animal embryogenesis. We designed a set of strategies, methods and algorithms to "sequence" the cell lineage tree as a branching process annotated in space and time. Our goal is to fully reconstruct the dynamics of cell divisions and movements from time-lapse series of high-resolution optical sections obtained by multiphoton laser scanning microscopy throughout embryonic development of live animals. Embryomics allows the automated tracking of events such as cell division and cell death in live embryos and give us access to parameters such as the rate of cell proliferation in time and space.
Embryo staining and mounting:
Wild type (070418a) and Zoep (081018a) zebrafish embryos were injected at the one cell stage with 200 pg mcherry/H2B RNA and 200 pg eGFP-ras prepared from PCS2+ constructs. Injected embryos were raised at 28.5 deg C for the next 3 hours. Embryos were

mounted in a 3cm Petri dish with a glass slide bottom, sealing a hole of 0,5 mm at the Petri dish centre where a Teflon tore (ALPHAnov) with a hole of 780 microns received the dechorionated embryo. Embryo was maintained and properly oriented by infiltrating around the embryo 0.5% low melting point agarose (SIGMA) in embryo medium. Temperature control in the room ($22 \deg C$) insured about $26 \deg C$ under the objective and development slowly drifted from the standard $28.5 \deg C$ developmental table.

Image acquisition:

The volume was imaged with a Leica DM6000 upright microscope SP5 MLSM equipped with a 20x/0,95NA W deeping lens Olympus objective. Field size is 700x700 in x, y. Voxel size is 1.37 microns. Simultaneous excitation with 1030 nm and 980 nm femtosecond pulsed laser beams was obtained from a single source (Amplitude t pulse 20) with part of the beam modified through a photonic crystal fiber (Amplitude).

Variable Information:

```
(0-2)     Identity information for each cell
(3 - 4)   Neighbors were calculated by Voronoi tessellation.
(8 - 11)  Displacement was calculated by vector field from original image.
(12 - 32) Using membrane segmentation, extract from the shape information.
(32 - 46) Intensity information by matching membrane shape & original image.
(47 - 81) Neighbors correlation on Membrane information
(82 - 151)Same as Membrane for nuclei segmentation.
```

Workflow of image processing:

- Filtering: Oimage filtering of both channels membranes and nucleus was performed through the methodology described in (Drblikova 2008) (O.Drblikova, K.Mikula, Semi-implicit diamond-cell finite volume scheme for 3D nonlinear tensor diffusion in coherence enhancing image filtering, in Finite Volumes for Complex Applications V: Problems and Perspectives (Eds. R.Eymard, J.M.Herard), ISTE and WILEY, London, 2008, pp. 343-35)

- Nucleus centre detection was performed through the methodology described in (Frolkovic 2007) (Frolkovic, P., Mikula, K., Peyrieras, N., & Sarti, A. 2007. Counting Number of Cells and Cell Segmentation Using Advection-Diffusion Equations. KYBERNETIKAPRAHA.)

- Segmentation: of membranes and nuclei shape was performed through the methodology described in (Mikula 2008) for nuclei and (Luengo 2008) for membranes. (K.Mikula, N.Peyrieras, M.Remesikova, A.Sarti, 3D embryogenesis image segmentation by the generalized subjective surface method using the finite volume technique, in Finite Volumes for Complex Applications V: Problems and Perspectives (Eds. R.Eymard, J.M.Herard), ISTE and WILEY, London, 2008, pp. 585-592) (Luengo-Oroz, M, Duloquin, L, Castro, C, Savy, T, Faure, E, Lombardot, B, Bourgine, P, Peyriéras, N, & Santos, A. 2008. Can voronoi diagram model cell geometries in early sea-urchin embryogenesis ? Biomedical Imaging : From Nano to Macro.)

- Tracking of cells and detection of mitoses was performed with the methodology described in Melani, C, Peyrieras, N, Mikula, K, & Zanella, C. 2007. Cells tracking in a live zebrafish embryo. Engineering in Medicine and Biology Society, Jan. 2007.

- Manual mitosis annotation was performed by 3 differents biologist expert.

Data split:
Only manually annotated data were used. For the development dataset ZEBRA a subset of samples corresponding to cells before mitosis or not in mitosis was selected and the task was to separate these two types of samples. For the final evaluation set called "E", the samples not in mitosis and after mitosis were selected to create another binary classification problem. Hence the two datesets overlapped. The orders of the features and samples were randomized to make it difficult to identify the common examples.

9.5. Data statistics

See Tables 1 and 2.

Table 11: **Variables of Embryology datasets.** All variables are continuous.

Index	Min	Max	Name
0	0	319552192	Cell ID
1	0	4	Annotation value (target)
2	0	479	Cell Time
3	0	56	Number of neighbors
4	-1024	5.388671	Neighborhoods tensor Deformation
5	30.14	661.710022	X Coord
6	53.43	680.890015	Y Coord
7	5.48	139.740005	Z Coord
8	-21.9182	12.539	X Deplacement
9	-21.3561	20.1138	Y Deplacement
10	-10.0698	16.4175	Z Deplacement
11	0.009766	25.465019	Velocity
12	0.004082	29.3297	Membrane Distance Center Gravity
13	0.000102	170803	Membrane Volume Segmentation
14	0	49407.39844	Membrane Volume Pixel
15	0.012692	5564.97998	Membrane Surface Area
16	0.229433	6.81659	Membrane Normalize Shape Index
17	22.3713	483.334015	Membrane X Gravity Center
18	39.674	498.429993	Membrane Y Gravity Center
19	5.2	101.963997	Membrane Z Gravity Center
20	0	0.663469	Membrane X Ellipse Axes Length
21	0	0.68352	Membrane Y Ellipse Axes Length
22	0	0.483781	Membrane Z Ellipse Axes Length
23	0.052632	inf	Membrane Elipse Ratio Elongation (Axes Max/Min)
24	0.57735	1	Membrane 0.0 Eigen Vectors Covariance Matrix
25	-0.707107	0.707107	Membrane 0.1 Eigen Vectors Covariance Matrix
26	-0.707107	0.707107	Membrane 0.2 Eigen Vectors Covariance Matrix
27	-0.815555	0.814614	Membrane 1.0 Eigen Vectors Covariance Matrix
28	0.412937	1	Membrane 1.1 Eigen Vectors Covariance Matrix
29	-0.707107	0.707107	Membrane 1.2 Eigen Vectors Covariance Matrix
30	-0.815264	0.814966	Membrane 2.0 Eigen Vectors Covariance Matrix
31	-0.827959	0.828677	Membrane 2.1 Eigen Vectors Covariance Matrix
32	0.334825	1	Membrane 2.2 Eigen Vectors Covariance Matrix
33	0	240.395004	Membrane Mean Intensity
34	0	255	Membrane Max Intensity

Continued on next page

Table 11 – continued from previous page

Index	Min	Max	Name
35	0	255	Membrane Min Intensity
36	0	82.144897	Membrane Sigma Intensity
37	0	2096780	Membrane Sum Intensity
38	0	230.858994	Membrane Mean Countour Intensity
39	0	255	Membrane Max Countour Intensity
40	0	255	Membrane Min Countour Intensity
41	0	88.0933	Membrane Sigma Countour Intensity
42	0	751864	Membrane Mean Sphere Intensity
43	0	240.395004	Membrane Max Sphere Intensity
44	0	255	Membrane Min Sphere Intensity
45	0	255	Membrane Sigma Sphere Intensity
46	0	82.144897	Membrane Sum Sphere Intensity
47	0	2096780	Neigbhors Membrane Distance Center Gravity
48	0.037032	12.6397	Neigbhors Membrane Volume Segmentation
49	1.13492	6568.560059	Neigbhors Membrane Volume Pixel
50	4.835757	6605.069824	Neigbhors Membrane Surface Area
51	8.099204	2495.22998	Neigbhors Membrane Normalize Shape Index
52	0.019907	1.50983	Neigbhors Membrane X Gravity Center
53	1.336024	467.737	Neigbhors Membrane Y Gravity Center
54	1.74803	475.114014	Neigbhors Membrane Z Gravity Center
55	0.297983	98.843803	Neigbhors Membrane X Ellipse Axes Length
56	0	0.228821	Neigbhors Membrane Y Ellipse Axes Length
57	0	0.197398	Neigbhors Membrane Z Ellipse Axes Length
58	0	0.133357	Neigbhors Membrane Elipse Ratio Elongation (Axes Max/Min)
59	0.056051	inf	Neigbhors Membrane 0.0 Eigen Vectors Covariance Matrix
60	0.015047	1	Neigbhors Membrane 0.1 Eigen Vectors Covariance Matrix
61	-0.630138	0.707107	Neigbhors Membrane 0.2 Eigen Vectors Covariance Matrix
62	-0.67426	0.678594	Neigbhors Membrane 1.0 Eigen Vectors Covariance Matrix
63	-0.708636	0.730605	Neigbhors Membrane 1.1 Eigen Vectors Covariance Matrix
64	0.014976	1	Neigbhors Membrane 1.2 Eigen Vectors Covariance Matrix
65	-0.698228	0.683967	Neigbhors Membrane 2.0 Eigen Vectors Covariance Matrix
66	-0.715458	0.727886	Neigbhors Membrane 2.1 Eigen Vectors Covariance Matrix
67	-0.766683	0.756487	Neigbhors Membrane 2.2 Eigen Vectors Covariance Matrix
68	0.01482	1	Neigbhors Membrane Mean Intensity
69	0.951452	87.638397	Neigbhors Membrane Max Intensity
70	2.933373	214	Neigbhors Membrane Min Intensity
71	0	29.444445	Neigbhors Membrane Sigma Intensity
72	0.449344	40.044201	Neigbhors Membrane Sum Intensity
73	391	228298	Neigbhors Membrane Mean Countour Intensity
74	0.843844	58.7286	Neigbhors Membrane Max Countour Intensity
75	3.26108	201	Neigbhors Membrane Min Countour Intensity
76	0	28.888887	Neigbhors Membrane Sigma Countour Intensity
77	0.476707	40.816002	Neigbhors Membrane Mean Sphere Intensity
78	1652.042236	177889	Neigbhors Membrane Max Sphere Intensity
79	0.951452	87.638397	Neigbhors Membrane Min Sphere Intensity
80	2.933373	214	Neigbhors Membrane Sigma Sphere Intensity
81	0	29.444445	Neigbhors Membrane Sum Sphere Intensity
82	0.449344	40.044201	Nucleus Distance Center Gravity
83	391	228298	Nucleus Volume Segmentation
84	0.004082	29.3297	Nucleus Volume Pixel

Continued on next page

Table 11 – continued from previous page

Index	Min	Max	Name
85	0.000102	170803	Nucleus Surface Area
86	0	49407.39844	Nucleus Normalize Shape Index
87	0.012692	5564.97998	Nucleus X Gravity Center
88	0.229433	6.81659	Nucleus Y Gravity Center
89	22.3713	483.334015	Nucleus Z Gravity Center
90	39.674	498.429993	Nucleus X Ellipse Axes Length
91	5.2	101.963997	Nucleus Y Ellipse Axes Length
92	0	0.663469	Nucleus Z Ellipse Axes Length
93	0	0.68352	Nucleus Elipse Ratio Elongation (Axes Max/Min)
94	0	0.483781	Nucleus 0.0 Eigen Vectors Covariance Matrix
95	0.052632	inf	Nucleus 0.1 Eigen Vectors Covariance Matrix
96	0.57735	1	Nucleus 0.2 Eigen Vectors Covariance Matrix
97	-0.707107	0.707107	Nucleus 1.0 Eigen Vectors Covariance Matrix
98	-0.707107	0.707107	Nucleus 1.1 Eigen Vectors Covariance Matrix
99	-0.815555	0.814614	Nucleus 1.2 Eigen Vectors Covariance Matrix
100	0.412937	1	Nucleus 2.0 Eigen Vectors Covariance Matrix
101	-0.707107	0.707107	Nucleus 2.1 Eigen Vectors Covariance Matrix
102	-0.815264	0.814966	Nucleus 2.2 Eigen Vectors Covariance Matrix
103	-0.827959	0.828677	Nucleus Mean Intensity
104	0.334825	1	Nucleus Max Intensity
105	0	240.395004	Nucleus Min Intensity
106	0	255	Nucleus Sigma Intensity
107	0	255	Nucleus Sum Intensity
108	0	82.144897	Nucleus Mean Countour Intensity
109	0	2096780	Nucleus Max Countour Intensity
110	0	230.858994	Nucleus Min Countour Intensity
111	0	255	Nucleus Sigma Countour Intensity
112	0	255	Nucleus Mean Sphere Intensity
113	0	88.0933	Nucleus Max Sphere Intensity
114	0	751864	Nucleus Min Sphere Intensity
115	0	240.395004	Nucleus Sigma Sphere Intensity
116	0	255	Nucleus Sum Sphere Intensity
117	0	255	Neigbhors Nucleus Distance Center Gravity
118	0	82.144897	Neigbhors Nucleus Volume Segmentation
119	0	2096780	Neigbhors Nucleus Volume Pixel
120	0.017163	21.813246	Neigbhors Nucleus Surface Area
121	0.027127	4100.373535	Neigbhors Nucleus Normalize Shape Index
122	0	356.075928	Neigbhors Nucleus X Gravity Center
123	0.209252	112.432922	Neigbhors Nucleus Y Gravity Center
124	0.019737	inf	Neigbhors Nucleus Z Gravity Center
125	1.204451	94.037766	Neigbhors Nucleus X Ellipse Axes Length
126	1.509158	122.508255	Neigbhors Nucleus Y Ellipse Axes Length
127	0.285211	24.912937	Neigbhors Nucleus Z Ellipse Axes Length
128	0	0.126924	Neigbhors Nucleus Elipse Ratio Elongation (Axes Max/Min)
129	0	0.099319	Neigbhors Nucleus 0.0 Eigen Vectors Covariance Matrix
130	0	0.042802	Neigbhors Nucleus 0.1 Eigen Vectors Covariance Matrix
131	0.004049	inf	Neigbhors Nucleus 0.2 Eigen Vectors Covariance Matrix
132	0.015451	0.210304	Neigbhors Nucleus 1.0 Eigen Vectors Covariance Matrix
133	-0.108076	0.121884	Neigbhors Nucleus 1.1 Eigen Vectors Covariance Matrix
134	-0.073733	0.068335	Neigbhors Nucleus 1.2 Eigen Vectors Covariance Matrix

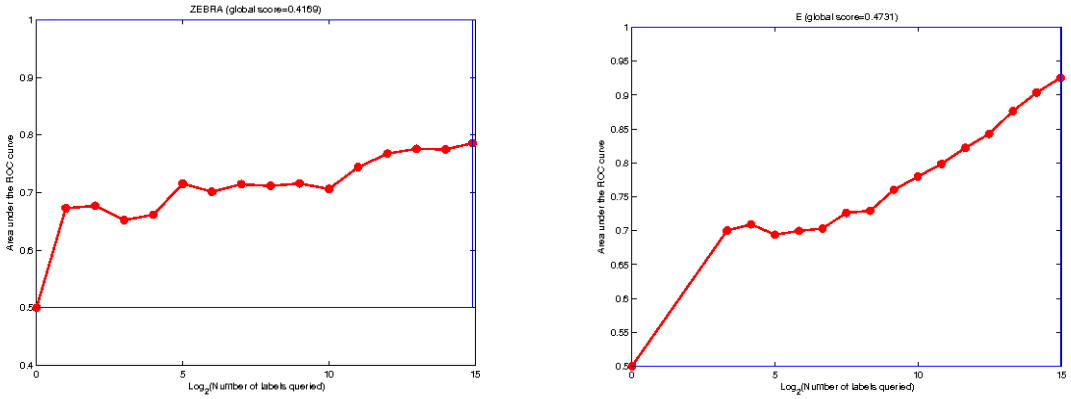
Continued on next page

Table 11 – continued from previous page

Index	Min	Max	Name
135	-0.121965	0.105965	Neigbhors Nucleus 2.0 Eigen Vectors Covariance Matrix
136	0.015619	0.21639	Neigbhors Nucleus 2.1 Eigen Vectors Covariance Matrix
137	-0.077092	0.083046	Neigbhors Nucleus 2.2 Eigen Vectors Covariance Matrix
138	-0.06829	0.079347	Neigbhors Nucleus Mean Intensity
139	-0.084374	0.084188	Neigbhors Nucleus Max Intensity
140	0.015546	0.246692	Neigbhors Nucleus Min Intensity
141	0	24.37923	Neigbhors Nucleus Sigma Intensity
142	0	54.3125	Neigbhors Nucleus Sum Intensity
143	0.034332	42.5	Neigbhors Nucleus Mean Countour Intensity
144	0	12.743444	Neigbhors Nucleus Max Countour Intensity
145	0	11538	Neigbhors Nucleus Min Countour Intensity
146	0	18.429482	Neigbhors Nucleus Sigma Countour Intensity
147	0	55.1875	Neigbhors Nucleus Mean Sphere Intensity
148	0.020833	42.5	Neigbhors Nucleus Max Sphere Intensity
149	0	12.822519	Neigbhors Nucleus Min Sphere Intensity
150	0	16145.33301	Neigbhors Nucleus Sigma Sphere Intensity
151	0	24.37923	Neigbhors Nucleus Sum Sphere Intensity

9.6. Baseline results

We show below a reference learning curve obtained for ZEBRA by the organizers and the learning curve of the overall winners for dataset E. More results are found on the webpage: `http://www.causality.inf.ethz.ch/activelearning.php?page=results#cont`.



(a) (b)

Figure 8: Reference results by Gavin Cawley for Zebra (a) and Results of IdealAnalytics, Intel on dataset E (b).