Design of experiments for the NIPS 2003 variable selection benchmark

Isabelle Guyon – July 2003
isabelle@clopinet.com

Background:

Results published in the field of feature or variable selection (see e.g. the special issue of JMLR on variable and feature selection: http://www.jmlr.org/papers/special/feature.html) are for the most part on different data sets or used different data splits, which make them hard to compare. We formatted a number of datasets for the purpose of benchmarking variable selection algorithms in a controlled manner[1]. The data sets were chosen to span a variety of domains (cancer prediction from mass-spectrometry data, handwritten digit recognition, text classification, and prediction of molecular activity). One dataset is artificial. We chose data sets that had sufficiently many examples to create a large enough test set to obtain statistically significant results. The input variables are continuous or binary, sparse or dense. All problems are two-class classification problems. The similarity of the tasks allows participants to enter results on all data sets. Other problems will be added in the future.

Method:

Preparing the data included the following steps:

- Preprocessing data to obtain features in the same numerical range (0 to 999 for continuous data and 0/1 for binary data).
- Adding "random" features distributed similarly to the real features. In what follows we refer to such features as **probes** to distinguish them from the real features. This will allow us to rank algorithms according to their ability to filter out irrelevant features.
- Randomizing the order of the patterns and the features to homogenize the data.
- Training and testing on various data splits using simple feature selection and classification methods to obtain baseline performances.
- Determining the approximate number of *test examples* needed for the test set to obtain statistically significant benchmark results using the rule-of-thumb $n_{test} = 100/p$, where p is the test set error rate (see What size test set gives good error rate estimates? *I. Guyon, J. Makhoul, R. Schwartz, and V. Vapnik*. PAMI, 20 (1), pages 52--64, IEEE. 1998, http://www.clopinet.com/isabelle/Papers/test-size.ps.Z). Since the test error rate of the classifiers of the benchmark is unknown, we used the results of the baseline method and added a few more examples.
- Splitting the data into training, validation and test set. The size of the validation set is usually smaller than that of the test set to keep as much training data as possible.

Both validation and test set truth-values (labels) are withheld during the benchmark. The validation set serves as development test set. During the time allotted to the participants to try methods on the data, participants are allowed to send the *validation set results* (in

[1] In this document, we do not make a distinction between features and variables. The benchmark addresses the problem of selecting input variables. Those may actually be features derived from the original variables using a preprocessing.

the form of classifier outputs) and obtain result scores. Such score are made available to all participants to stimulate research. At the end of the benchmark, the participants send their *test set results*. The scores on the test set results are disclosed simultaneously to all participants after the benchmark is over.

Data formats:
All the data sets are in the same format and include 8 files in ASCII format:
dataname.param: Parameters and statistics about the data
dataname.feat: Identities of the features (in the order the features are found in the data).
dataname_train.data: Training set (a spase or a regular matrix, patterns in lines, features in columns).
dataname_valid.data: Validation set.
dataname_test.data: Test set.
dataname_train.labels: Labels (truth values of the classes) for training examples.
dataname_valid.labels: Validation set labels (withheld during the benchmark).
dataname_test.labels: Test set labels (withheld during the benchmark).
The matrix data formats used are:

- For regular matrices: a space delimited file with a new-line character at the end of each line.
- For sparse matrices with binary values: for each line of the matrix, a space delimited list of indices of the non-zero values. A new-line character at the end of each line.
- For sparse matrices with non-binary values: for each line of the matrix, a space delimited list of indices of the non-zero values followed by the value itself, separated from it index by a colon. A new-line character at the end of each line.

The results on each dataset should be formatted in 7 ASCII files:
dataname_train.resu: +-1 classifier outputs for training examples (mandatory for **final** submissions).
dataname_valid.resu: +-1 classifier outputs for validation examples (mandatory for **development and final** submissions).
dataname_test.resu: +-1 classifier outputs for test examples (mandatory for **final** submissions).
dataname_train.conf: confidence values for training examples (optional).
dataname_valid.conf: confidence values for validation examples (optional).
dataname_test.conf: confidence values for test examples (optional).
dataname.feat: list of features selected (one integer feature number per line, starting from one, ordered from the most important to the least important if such order exists). If no list of features is provided, it will be assumed that all the features were used.
Format for classifier outputs:
- All .resu files should have one +-1 integer value per line indicating the prediction for the various patterns.
- All .conf files should have one decimal positive numeric value per line indicating classification confidence. The confidence values can be the absolute discriminant values. They do not need to be normalized to look like probabilities. They will be used to compute ROC curves and Area Under such Curve (AUC).

Result rating:
The classification results are rated with the balanced error rate (the average of the error rate on training examples and on test examples). The area under the ROC curve is also be computed, if the participants provide classification confidence scores in addition to class label predictions. But **the relative strength of classifiers is judged only on the balanced error rate**. The participants are invited to provide the list of features used. **For methods having performance differences that are not statistically significant, the method using the smallest number of features wins**. If no feature set is provided, it is assumed that all the features were used. The organizers may then provide the participants with one or several test sets containing only the features selected to verify the accuracy of the classifier when it uses those features only. The proportion of random probes in the feature set is also be computed. It is used to assess the relative strength of method with non-statistically significantly different error rates and a relative difference in number of features that is less than 5%. **In that case, the method with smallest number of random probes in the feature set wins**.

Dataset A: ARCENE

1) Topic

The task of ARCENE is to distinguish **cancer** *versus* normal patterns from mass-spectrometric data. This is a two-class classification problem with continuous input variables.

2) Sources

a. Original owners

The data were obtained from two sources: The National Cancer Institute (NCI) and the Eastern Virginia Medical School (EVMS). All the data consist of mass-spectra obtained with the SELDI technique. The samples include patients with cancer (ovarian or prostate cancer), and healthy or control patients.

NCI ovarian data:
The data were originally obtained from http://clinicalproteomics.steem.com/download-ovar.php. We use the 8/7/02 data set:
http://clinicalproteomics.steem.com/Ovarian%20Dataset%208-7-02.zip.
The data includes 253 spectra, including 91 controls and 162 cancer spectra.
Number of features: 15154.

NCI prostate cancer data:
The data were originally obtained from
http://clinicalproteomics.steem.com/JNCI%20Data%207-3-02.zip on the web page
http://clinicalproteomics.steem.com/download-prost.php.
There are a total of 322 samples: 63 samples with no evidence of disease and PSA level less than 1; 190 samples with benign prostate with PSA levels greater than 4; 26 samples with prostate cancer with PSA levels 4 through 10; 43 samples with prostate cancer with PSA levels greater than 10. Therefore, there are 253 normal samples and 69 disease samples. The original training set is composed of 56 samples:

- 25 samples with no evidence of disease and PSA level less than 1 ng/ml.
- 31 biopsy-proven prostate cancer with PSA level larger than 4 ng/ml.

But the exact split is not given in the paper or on the web site. The original test set contains the remaining 266 samples (38 cancer and 228 normal).
Number of features: 15154.

EVMS prostate cancer data:
The data is downloadable from:
http://www.evms.edu/vpc/seldi/.
The training data data includes 652 spectra from 326 patients (spectra are in duplicate) and includes 318 controls and 334 cancer spectra. Study population: 167 prostate cancer (84 state 1 and 2; 83 stage 3 and 4), 77 benign prostate hyperplasia, and 82 age-matched normals. The test data includes 60 additional patients. The labels for the test set are not provided with the data, so the test spectra are not used for the benchmark.
Number of features: 48538.

b. Donor of database

This version of the database was prepared for the NIPS 2003 variable and feature selection benchmark by Isabelle Guyon, 955 Creston Road, Berkeley, CA 94708, USA (isabelle@clopinet.com).

c. Date received: August 2003.

3) Past usage

NCI ovarian cancer original paper:
"Use of proteomic patterns in serum to identify ovarian cancer *Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta.* THE LANCET • Vol 359 • February 16, 2002 • www.thelancet.com" are so far not reproducible.
Note: The data used is a newer set of spectra obtained after the publication of the paper and of better quality.
100% accuracy is easily achieved on the test set using various data splits on this version of the data.
NCI prostate cancer original paper:
Serum proteomic patterns for detection of prostate cancer. Petricoin et al. Journal of the NCI, Vol. 94, No. 20, Oct. 16, 2002. The test results of the paper are shown in Table A.1.

FP	FN	TP	TN	Error	1-error	Specificity	Sensitivity
51	2	36	177	20.30%	79.70%	77.63%	94.74%

Table A.1: Results of Petricoin et al. on the NCI prostate cancer data. Fp=false positive, FN=false negative, TP=true positive, TN=true negative.
Error=(FP+FN)/(FP+FN+TP+TN), Specificity=TN/(TN+FP), Sensitivity=TP/(TP+FN).

EVMS prostate cancer original paper:
Serum Protein Fingerprinting Coupled with a Pattern-matching Algorithm Distinguishes Prostate Cancer from Benign Prostate Hyperplasia and Healthy Men, Bao-Ling Adam, et al., CANCER RESEARCH 62, 3609–3614, July 1, 2002.

In the following excerpt from the original paper some baseline results are reported:
"Surface enhanced laser desorption/ionization mass spectrometry protein profiles of serum from 167 PCA patients, 77 patients with benign prostate hyperplasia, and 82 age-matched unaffected healthy men were used to train and develop a decision tree classification algorithm that used a nine-protein mass pattern that correctly classified 96% of the samples. A blinded test set, separated from the training set by a stratified random sampling before the analysis, was used to determine the sensitivity and specificity of the classification system. A sensitivity of 83%, a specificity of 97%, and a positive predictive value of 96% for the study population and 91% for the general population were obtained when comparing the PCA *versus* noncancer (benign prostate hyperplasia/healthy men) groups**."**

4) Experimental design

We merge the datasets from the three different sources (253+322+326=901 samples). We obtained 91+253+159=503 control samples (negative class) and 162+69+167=398 cancer samples (positive class). The motivations for merging datasets include:

- Obtaining enough data to be able to cut a sufficient size test set.
- Creating a problem where possibly non-linear classifiers and non-linear feature selection methods might outperform linear methods. The reason is that there will be in each class different clusters corresponding differences in disease, gender, and sample preparation.
- Finding out whether there are features that are generic of the separation cancer *vs*. normal across various cancers.

We designed a preprocessing that is suitable for mass-spec data and applied it to all the data sets to reduce the disparity between data sources. The preprocessing consists of the following steps:

- **Limiting the mass range:** We eliminated small masses under m/z=200 that include usually chemical noise specific to the MALDI/SELDI process (influence of the "matrix"). We also eliminated large masses over m/z=10000 because few features are usually relevant in that domain and we needed to compress the data.
- **Averaging the technical repeats:** In the EVMS data, two technical repeats were available. We averaged them because we wanted to have examples in the test set that are independent so that we can apply simple statistical tests.
- **Removing the baseline:** We subtracted in a window the median of the 20% smallest values. An example of baseline detection is shown in Figure A.1.
- **Smoothing:** The spectra were slightly smoothed with an exponential kernel in a window of size 9.
- **Re-scaling**: The spectra were divided by the median of the 5% top values.
- **Taking the square root**. The square root of the all values was taken.
- **Aligning the spectra**: We slightly shifted the spectra collections of the three datasets so that the peaks of the average spectrum would be better aligned (Figures A.2 and A.3). As a result, the mass-over-charge (m/z) values that identify the features in the aligned data are imprecise. We took the NCI prostate cancer m/z as reference.
- **Limiting more the mass range**: To eliminate border effects, the spectra border were cut.

- **Soft thresholding the values:** After examining the distribution of values in the data matrix, we subtracted a threshold and equaled to zero all the resulting values that were negative. In this way, we kept only about 50% of non-zero value, which represents significant data compression (see Figure A.4).
- **Quantizing:** We quantized the values to 1000 levels.

The resulting data set including all training and test data merged from the three sources has 901 patterns from 2 classes and 9500 features. We remove one pattern to obtain the round number 900. At every step, we checked that the change in performance of a linear SVM classifier trained and tested on a random split of the data was not significant. On that basis, we have some confidence that our preprocessing did not alter significantly the information content of the data. We further manipulated the data to add random "probes":

- We identified the region of the spectra with least information content using an interval search for the region that gave worst prediction performance of a linear SVM (indices 2250-4750). We replaced the features in that region by "random probes" obtained by randomly permuting the values in the columns of the data matrix.
- We identified another region of low information content: 6500-7000. We added 500 random probes that are permutations of those features.

After such manipulations, the data had 10000 features, including 7000 real features and 3000 random probes. The reason for not adding more probes is purely practical: non-sparse data cannot be compressed sufficiently to be stored and transferred easily in the context of a benchmark.

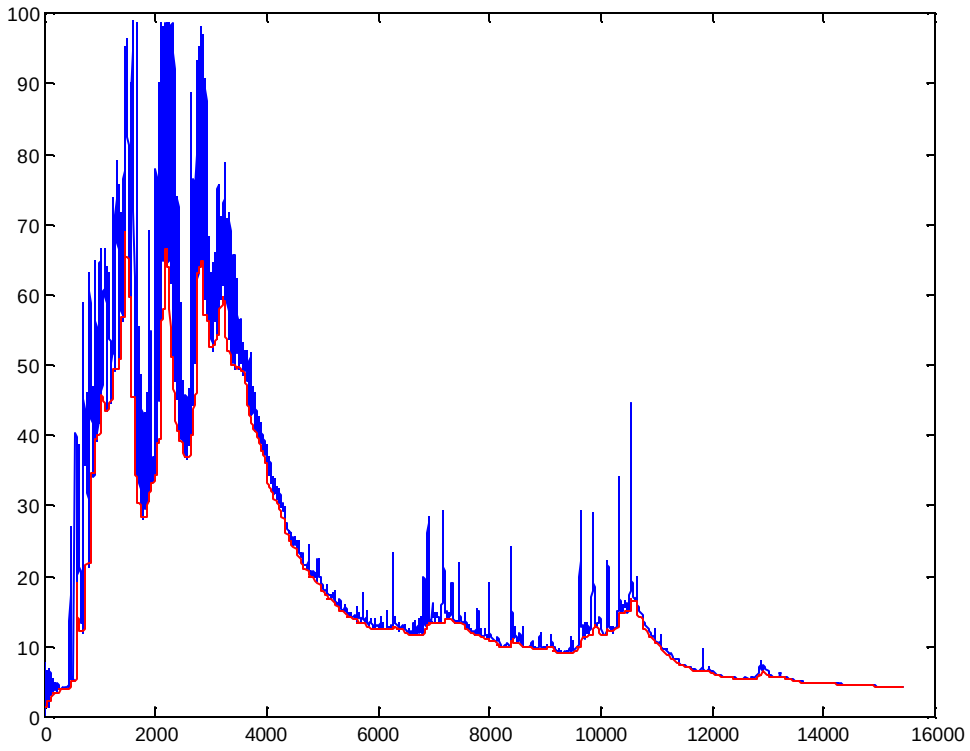

Figure A.1: Example of baseline detection (EVMS data).

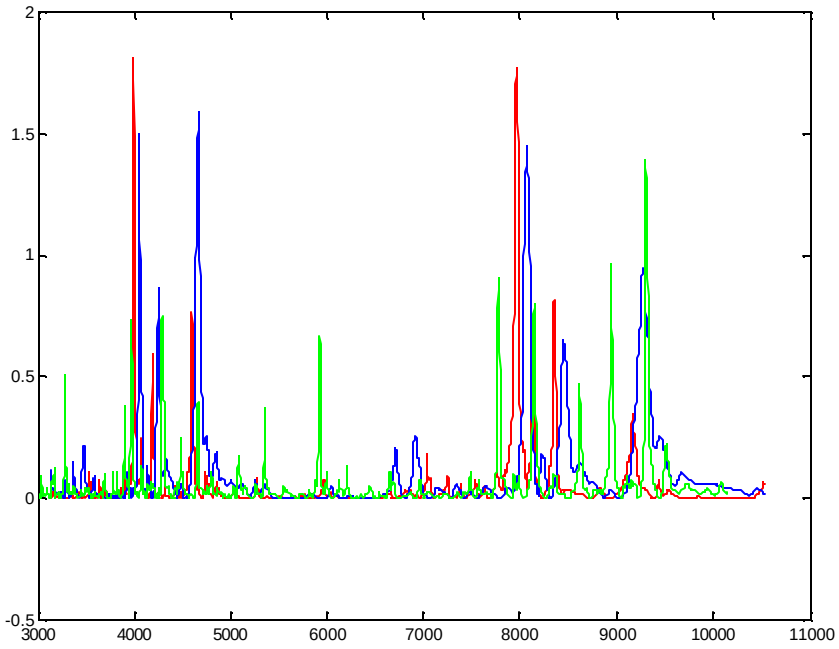

Figure A.2: Central part of the spectra before alignment. We show in red the average NCI ovarian spectra, in blue the average NCI prostate spectra, and in green the average EVMS prostate spectra.

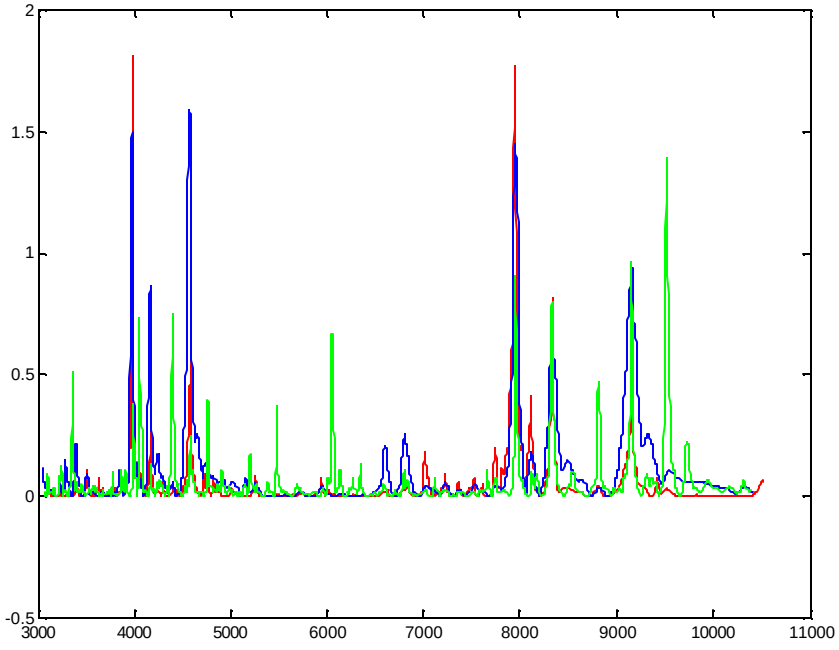

Figure A.3: Central part of the spectra after alignment. We show in red the average NCI ovarian spectra, in blue the average NCI prostate spectra, and in green the average EVMS prostate spectra.

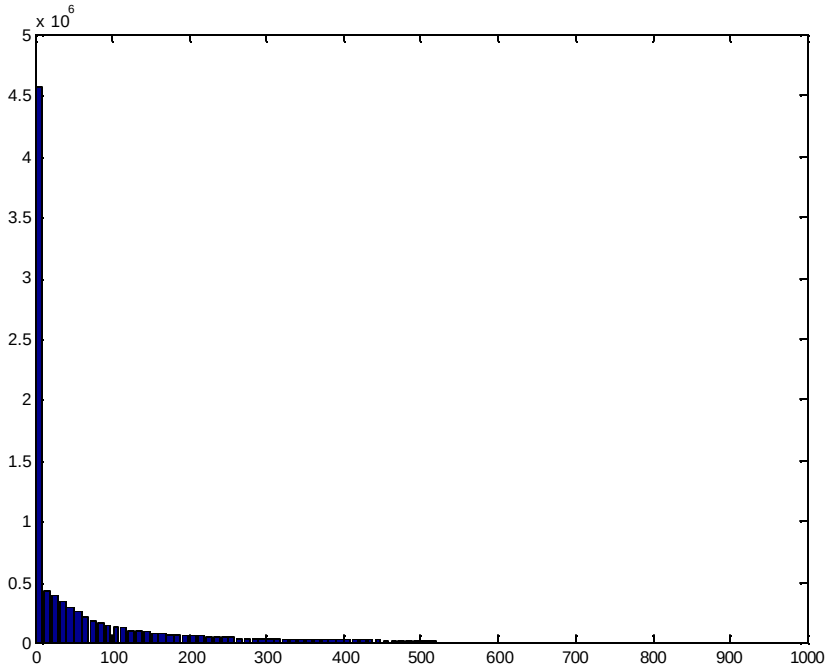

Figure A.4: Distributions of the values in the ARCENE data after preprocessing.

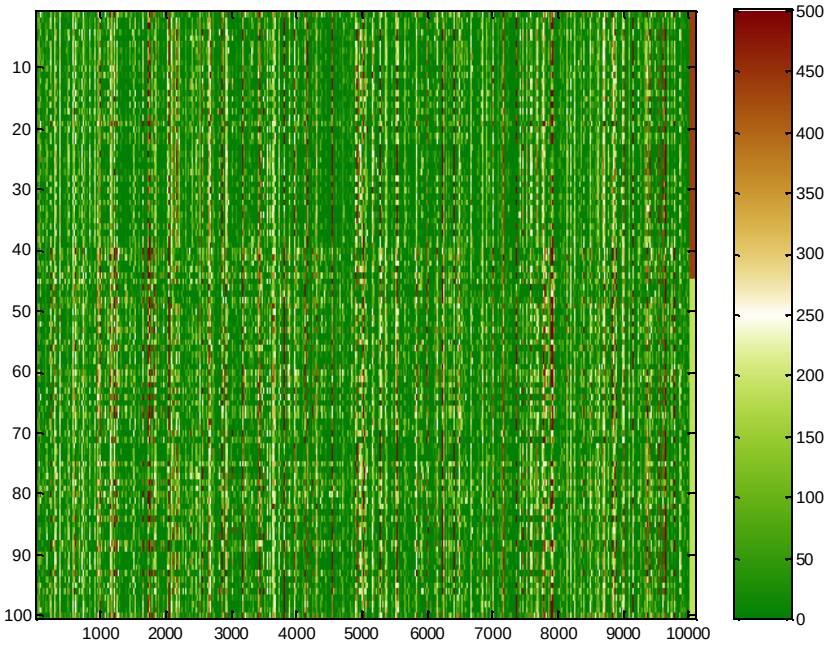

Figure A.5: Heat map of the training set of the ARCENE data. We represent the data matrix (patients in line and features in columns). The values are clipped at 500 to increase the contrast. The values are then mapped to colors according to the color-map on the right. The stripe beyond the 10000 feature index indicated the class labels: +1=red, -1=green.

5) Number of examples and class distribution

	Positive ex.	Negative ex.	Total	Check sum
Training set	44	56	100	70726744
Validation set	44	56	100	71410108
Test set	310	390	700	493023349
All	398	502	900	635160201

6) Type of input variables and variable statistics

Real variables	Random probes	Total
7000	3000	10000

All variables are **integer** quantized on 1000 levels. There are **no missing values**. The data is not very sparse, but for data compression reasons, we thresholded the values. Approximately 50% of the entries are non zero. The data was saved as a **non-sparse** matrix.

7) Results of the run of the lambda method and linear SVM

Before the benchmark, we ran some simple methods to determine what an appropriate number of examples should be. The "lambda" method (provided with the sample code) had approximately a 30% test error rate ans a linear SVM trained on all features a 15% error rate. The rule of thumb number_of_test_examples=100/test_errate=100/.15=667 led us to keep 700 examples for testing.
The best benchmark error rates are of the order 15%, which confirms that our estimate was correct.

Dataset B: GISETTE

1) Topic

The task of GISETTE is to discriminate between to confusable handwritten **digits**: the four and the nine. This is a two-class classification problem with sparse continuous input variables.

2) Sources

a. Original owners

The data set was constructed from the MNIST data that is made available by Yann LeCun of the NEC Research Institute at http://yann.lecun.com/exdb/mnist/.
The digits have been size-normalized and centered in a fixed-size image of dimension 28x28. We show examples of digits in Figure B1.