Rare category exploitation

Harel Lustiger

Dept. of Statistics and Operations Research, Tel-Aviv University

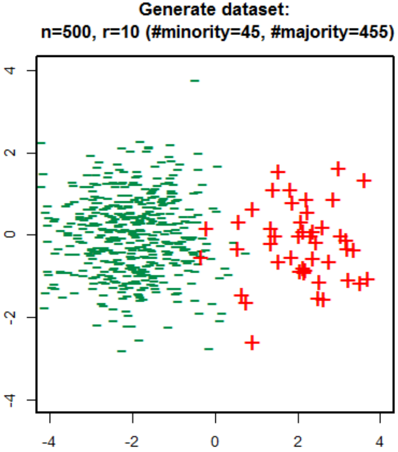
January 16, 2017

Goals

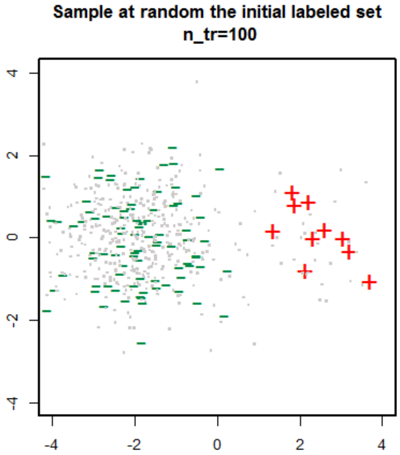
Quickly find the minority cases within an unlabeled data set given that:

1. The process can be carried out in a sequential fashion.
2. The sequence has a finite time horizon.

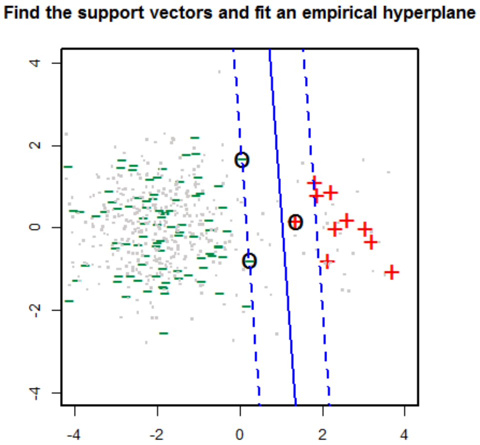
Toy Example: Scenario 1 (Non-linearly separable classes)



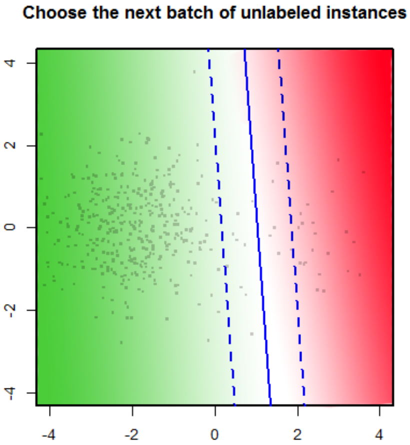
Toy Example: Scenario 1 (Non-linearly separable classes)



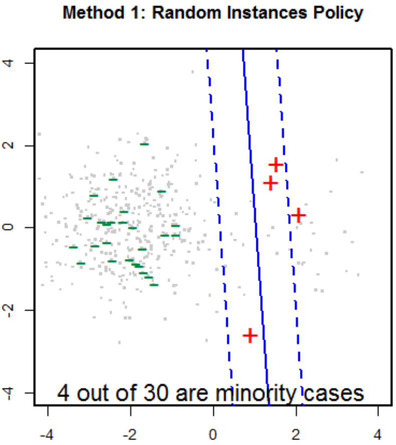
Toy Example: Scenario 1 (Non-linearly separable classes)



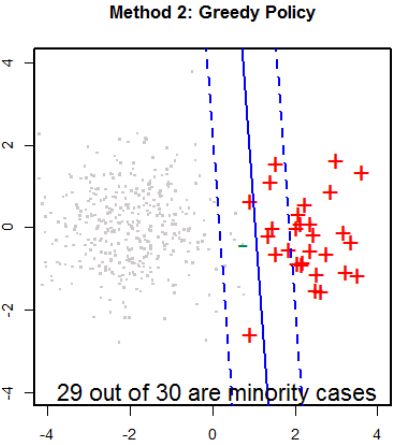
Toy Example: Scenario 1 (Non-linearly separable classes)



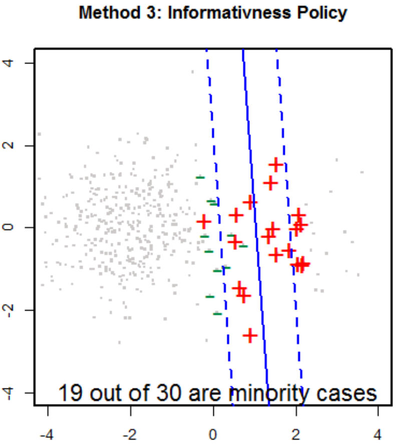
Toy Example: Scenario 1 (Non-linearly separable classes)



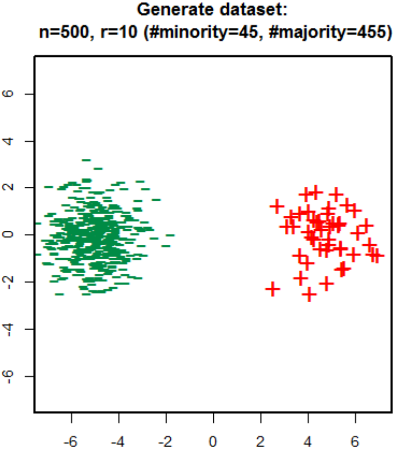
Toy Example: Scenario 1 (Non-linearly separable classes)



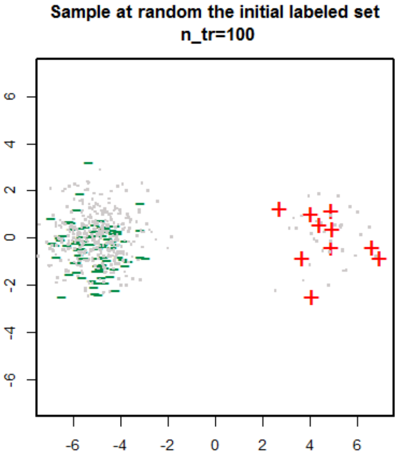
Toy Example: Scenario 1 (Non-linearly separable classes)



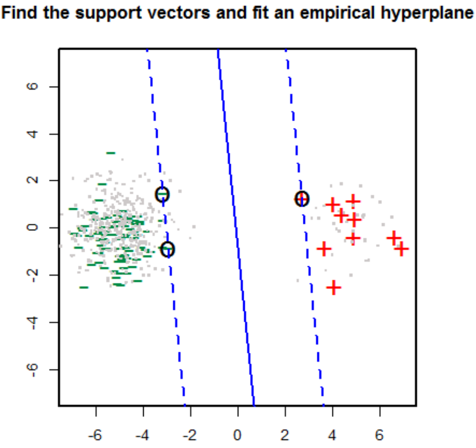
Toy Example: Scenario 2 (Highly separable classes)



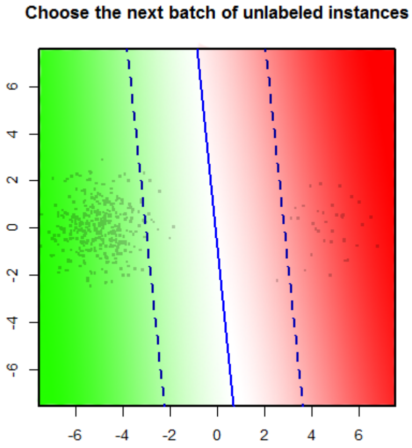
Toy Example: Scenario 2 (Highly separable classes)



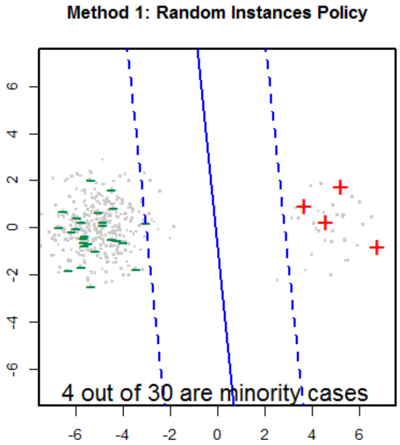
Toy Example: Scenario 2 (Highly separable classes)



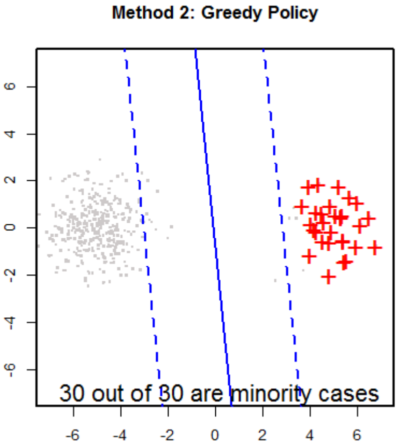
Toy Example: Scenario 2 (Highly separable classes)



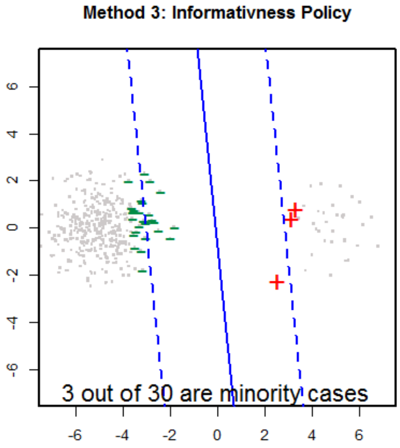
Toy Example: Scenario 2 (Highly separable classes)



Toy Example: Scenario 2 (Highly separable classes)



Toy Example: Scenario 2 (Highly separable classes)



Model Assessment and Selection

"The generalization performance of a learning method relates to its prediction capability on independent test data."[1]

While this is typically true, under our settings it is not the case:

1. We have no test set since our objective is to discover the minority cases within the unlabeled set.
2. When the experiment is terminated, the model becomes obsolete, such that no more unseen data is predicted.

[1] Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning.

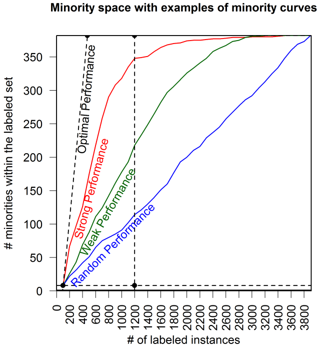
Evaluation Metrics: attributes

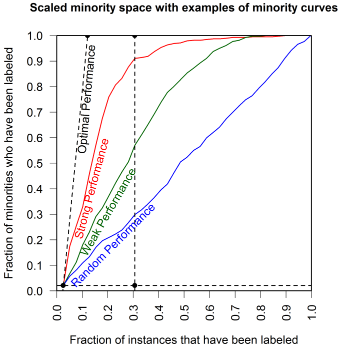
In the case of rare category exploitation, an adequate performance measure should be accountable for four properties of the scenario:

1. binary classification task
2. unbalanced data set
3. sequential experiment, and
4. finite horizon

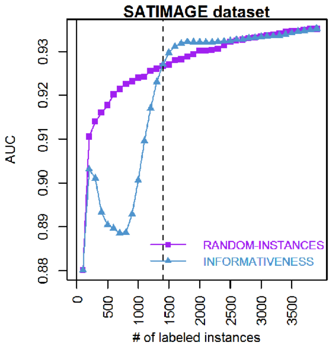
Evaluation Metrics: Temporal-Minority space

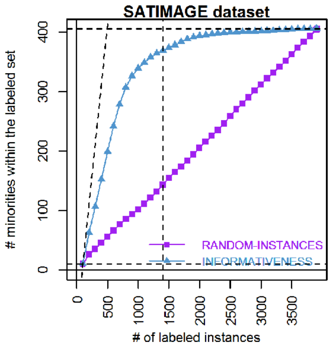
- In the continuum of the experiment, rare category exploitation is expressed in a confined subspace.
  - Its right boundary, here depicted by a dashed line at $x = 1200$ or $x \approx 0.3$ is driven by the problem constraints and expresses the finite horizon of the sequential experiment.





Evaluation Metrics: ROC space vs. TM space



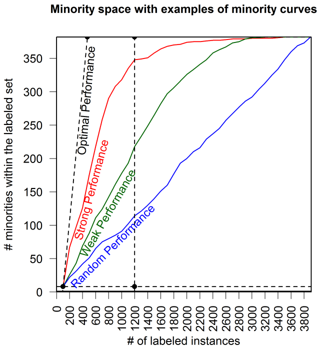


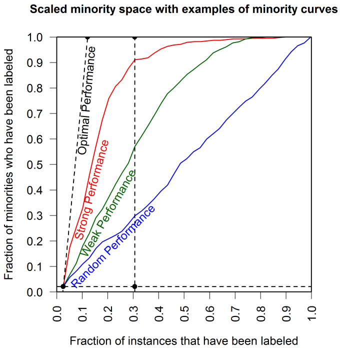
Evaluation Metrics: Comparing different policies

Results expressed in graphical measures can be more difficult to interpret than those reported in a single measure.

- the **total number of minority cases found at the end of the experiment**.
  - Because of its conciseness, summarizing only the endpoint into a scalar metric, it lacks informativeness.
  - It does not reflect the quickness property of the objective behind rare category exploitation.
- Instead, we propose to integrate the area under the curve in temporal-minority space **AUC-TM**.
  - Since the TM space is confined by a closed shape, we know the maximum achievable AUC-TM and therefore we can scale it such that AUC-TM $\in [0, 1]$

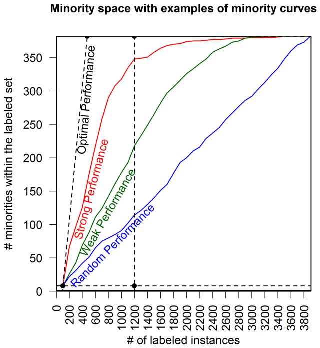
Evaluation Metrics: AUC-TM

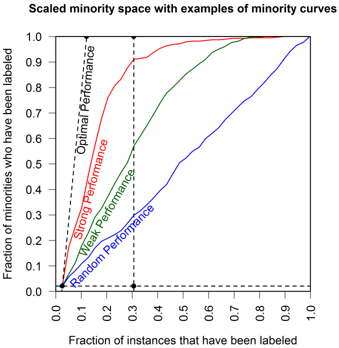




- If a policy has near `Optimal Performance`, it quickly detects most if not all the minority cases in the unlabeled set and the `AUC-TM` will be close to 1.
- Note: `AUC-TM` is not necessarily $\frac{1}{2}$ for `Random Performance`.

Evaluation Metrics: AUC−TM





- AUC−TM assigns a value to finding minority cases sooner rather than later.
- AUC−TM allows us to conclude that a policy is superior to a second policy if it dominates the other for most or all of the points along their TM curves.

Experiments: Setup

Datasets

1. ABALONE
2. LETTER
3. SETIMAGE

Algorithms

1. Random Instances policy
2. Greedy policy
3. Informativeness policy
4. Semi Uniform policy
5. ϵ-Greedy policy

Classifiers

1. SVM
2. Logistic regression

Experiments: Analysis of the experimental results

A policy is most useful if it exhibits robust performance across multiple settings, and does not perform poorly in any setting.

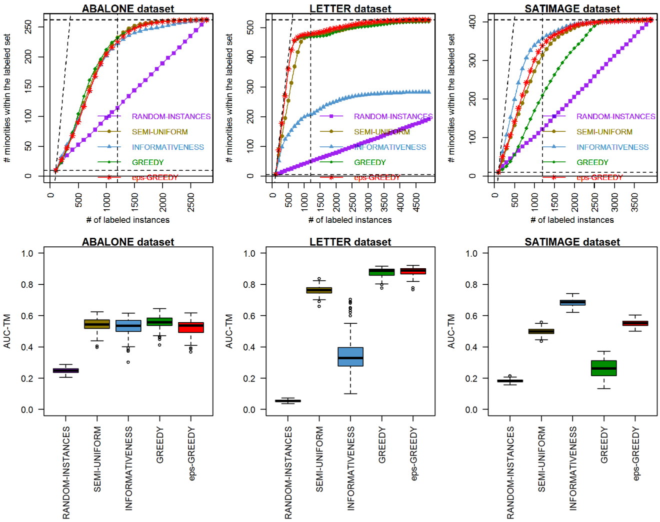
The ideal outcome of the experiment then, is a policy which satisfies two criteria:

1. **better** than the benchmark of non-data-driven policy, and
2. **not substantially worse** than other data-driven policies

To check these two assumptions, we perform Wilcoxon signed-rank test for the differences between policies' `AUC-TM` across 100 repetitions for one year of ongoing sequential experimentation (12 epochs).

With 4+12 tests, 3 data sets and 2 classifiers families, the p-value threshold placed at 5e-4 (0.05/96).

Experiments: Empirical Evaluation



Experiments: Results summary

	SEMIUNIFORM	INFORMATIVENESS	GREEDY	ϵ-GREEDY
# of rejections	9	11	3	4
Mean percentage difference	89.3	65.1	46.5	92.6

2

[2] out of 18 possible rejections