# Machine Learning Algorithms for Security and Image/Video Classification

## A Weighted Probabilistic Approach to the PU Learning Problem

Kristen Jaskie, Andreas Spanias
*SenSIP Center, School of ECEE, Arizona State University.*

SenSIP INDUSTRY CONSORTIUM

## MOTIVATION AND EXAMPLES

**Traditional classification requires well-labeled data.**

- **Both positive and negative labels**

**Negative data is expensive in many interesting problems:**

- **Ex: Satellite Image Object Classification**
  - **Known positive set:** Images with the desired object (Ex: New Military Installations in N. Korea, Archeological Sites, etc…)
  - **Unlabeled set:** All other images

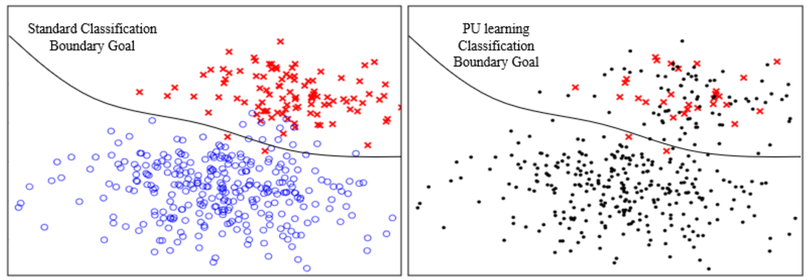
- **Ex: Cancer Gene Identification**
  - **Known positive set:** Genes known to influence cancer likelihood
  - **Unlabeled set:** All other genes in the genome
- **Ex: Abnormal ECG Signal Detection/Identification**
- **Ex: Fraud Detection**

**RESULT:**

- **Some Positive Labels**
- **No Negative Labels**

## GOAL

- Given data samples $x$ and data labels $y$
- We want to learn a probabilistic classifier $p(y = 1|x)$

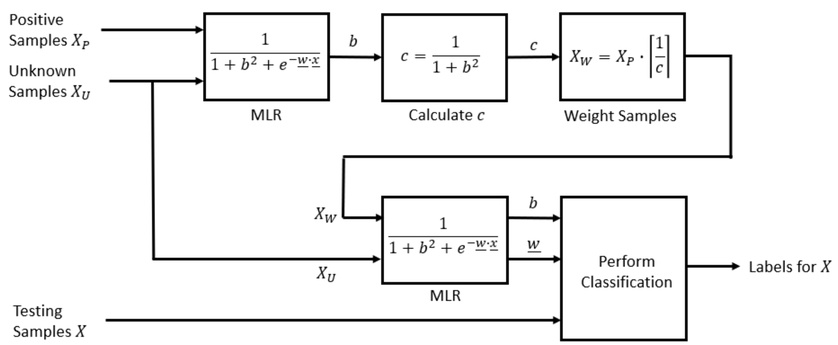



(a) Supervised classification learning problem with all labels known.

(b) Positive Unlabeled learning problem with only a percentage of positive labels known and all others unknown.

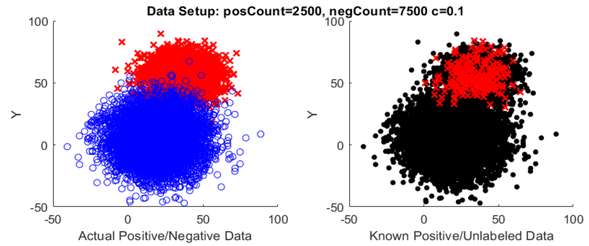
## WEIGHTED MODIFIED LOGISTIC REGRESSION (WMLR)

- Step 1: Use a Modified Logistic Regression (MLR) to calculate the probability $c$ that a positive sample is labeled positive.
- Step 2: Weight the labeled positive samples based on $c$
- Step 3: Use the MLR to learn a decision boundary to identify the positive and negative classes.




## RESULTS

- We compared our algorithm with the previous state of the art PU estimators from [1] and with standard logistic regression.
- F-Score metric shows significant and consistent improvement over previous algorithms.




**True c = 0.1**

| | Standard LR | Elkan e1 | Elkan e2 | Elkan e3 | Modified LR | WMLR |
|---|---|---|---|---|---|---|
| Confusion Matrix | 0 0 / 2500 7500 | 1922 10 / 578 7490 | 2038 16 / 462 7484 | 126 0 / 2374 7500 | 52 0 / 2448 7500 | 2262 36 / 238 7464 |
| C_hat | NA | 0.10297 | 0.093212 | 0.44351 | 0.35829 | |
| Accuracy | 0.75 | 0.9412 | 0.9522 | 0.7626 | 0.7552 | 0.9726 |
| Precision | NaN | 0.99482 | 0.99221 | 1 | 1 | 0.98433 |
| Recall | 0 | 0.7688 | 0.8152 | 0.0504 | 0.0208 | 0.9048 |
| F-score | NaN | 0.86733 | 0.89504 | 0.095963 | 0.040752 | 0.94289 |

**True c = 0.5**

| | Standard LR | Elkan e1 | Elkan e2 | Elkan e3 | Modified LR | WMLR |
|---|---|---|---|---|---|---|
| Confusion Matrix | 1043 0 / 1457 7500 | 2209 23 / 291 7477 | 2232 29 / 268 7471 | 1126 0 / 1374 7500 | 2331 48 / 169 7452 | 2424 91 / 76 7409 |
| C_hat | NA | 45528 | 0.43801 | 0.96163 | 0.53825 | |
| Accuracy | 0.8543 | 0.9686 | 0.9703 | 0.8626 | 0.9783 | 0.9833 |
| Precision | 1 | 0.9897 | 0.98717 | 1 | 0.97982 | 0.96382 |
| Recall | 0.4172 | 0.8836 | 0.8928 | 0.4504 | 0.9324 | 0.9696 |
| F-score | 0.58877 | 0.93364 | 0.93762 | 0.62107 | 0.95552 | 0.9667 |

**True c = 0.9**

| | Standard LR | Elkan e1 | Elkan e2 | Elkan e3 | Modified LR | WMLR |
|---|---|---|---|---|---|---|
| Confusion Matrix | 2293 44 / 207 7456 | 2359 66 / 141 7434 | 2358 66 / 142 7434 | 2293 44 / 207 7456 | 2396 81 / 104 7419 | 2442 124 / 58 7376 |
| C_hat | NA | 0.83075 | 0.83361 | 0.99977 | 0.91549 | |
| Accuracy | 0.9749 | 0.9793 | 0.9792 | 0.9749 | 0.9815 | 0.9818 |
| Precision | 0.98117 | 0.97278 | 0.97277 | 0.98117 | 0.9673 | 0.95168 |
| Recall | 0.9172 | 0.9436 | 0.9432 | 0.9172 | 0.9584 | 0.9768 |
| F-score | 0.94811 | 0.95797 | 0.95776 | 0.94811 | 0.96283 | 0.96407 |

## REFERENCES

[1] C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In Proceedings of the Fourteenth International Conference on Knowledge Discovery and Data Mining (KDD '08 ).