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

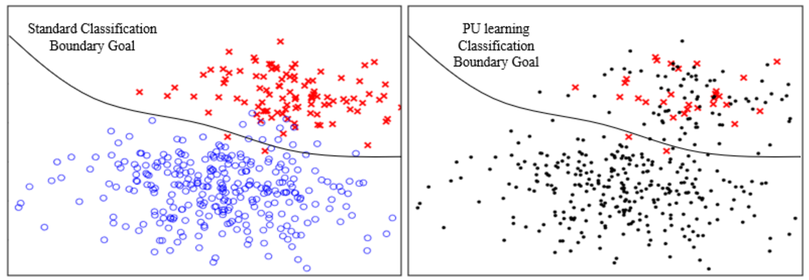
- **Ex: Cancer Gene Identification**
  - **Known positive set:** Genes known to influence cancer likelihood
  - **Unlabeled set:** All other genes in the genome
- **Ex: Abnormal ECG Signal Detection/Identification**
- **Ex: Fraud Detection**

RESULT:

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

GOAL

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

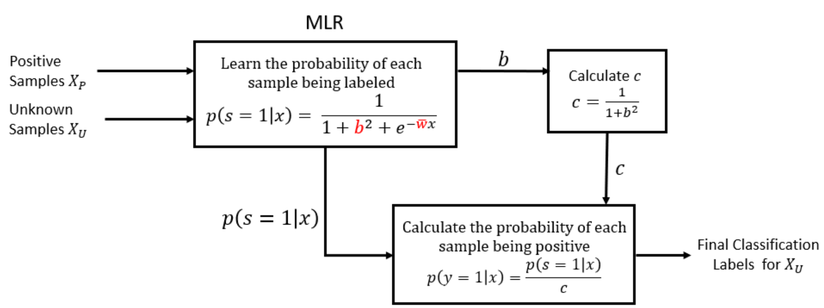


(a) Supervised classification learning problem with all labels known.

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

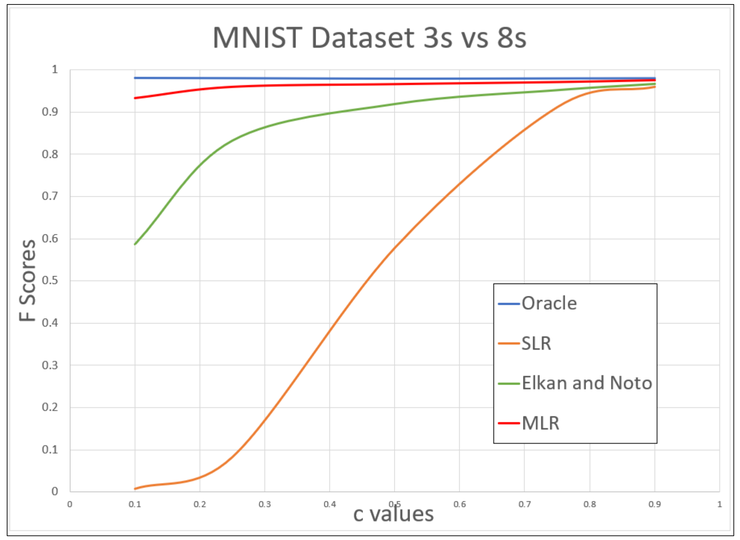
MODIFIED LOGISTIC REGRESSION (MLR)

- Use a Modified Logistic Regression (MLR) to learn a non-traditional classifier showing:
  - The probability of a sample being labeled $p(s = 1|x)$
  - The percentage c of true positive samples that are labelled positive
- Calculate the probability of a sample being positive using $p(s = 1|x)$ and c



RESULTS

- We compared our algorithm with estimators from [1] and with standard logistic regression.
- We used an F-Score metric for evaluative purposes to compensate for frequently uneven class sizes.
- The F-Score is the harmonic mean of the recall and precision and ranges from 0 to 1.
- F-Score metric shows significant and consistent improvement over previous algorithms.



REFERENCES

[1] C. Elkan and K. Noto, "Learning classifiers from only positive and unlabeled data," in *SIGKDD*, Las Vegas, ACM, pp. 213–20, Aug. 2008.

[2] K. Jaskie and A. S. Spanias, "Positive and Unlabeled Learning Algorithms and Applications : a Survey," in *IEEE IISA*, Patras, Greece, IEEE, pp. 1–8, Jul. 2019.

[3] K. Jaskie, C. Elkan, and A. Spanias, "A Modified Logistic Regression for Positive and Unlabeled Learning," in *IEEE Asilomar*, Pacific Grove, California, IEEE, Nov. 2019.