POSITIVE AND UNLABELED MACHINE LEARNING

Kristen Jaskie, SenSIP Center, School of ECEE, ASU

***Abstract* – The positive and unlabeled learning problem is a semi-supervised binary classification problem over a set of known positive samples, no known negative samples, and a large amount of unlabeled data. The probability of an unlabeled sample being positive is unknown. We build on previous work and introduce a new algorithm using a modified logistic regression algorithm to solve the problem.**

***Index Terms*—PU learning, positive unlabeled learning, machine learning, semi-supervised learning, classification**

1. PROJECT DESCRIPTION

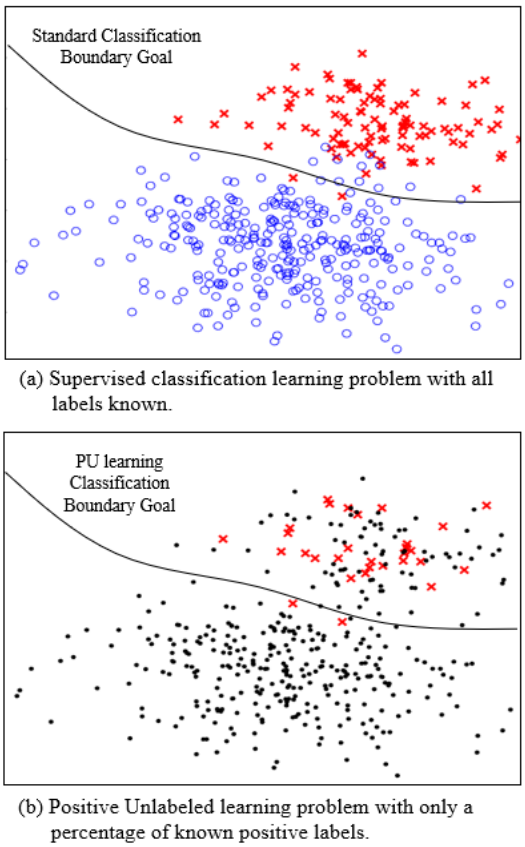
In supervised binary classification, training samples from two unknown data distributions, labeled as positive and negative, are available. The goal is to identify a classification or decision boundary that effectively separates these two classes. New data samples with unknown labels can then be identified as belonging to either the positive or negative classes [1], [2]. Support vector machines (SVMs), logistic regression, neural network learning algorithms, and others are used to solve this problem.

Real world data is often only partially labeled. Because completely labeling data can be expensive or even impossible, a common scenario involves having only a small number of labeled samples from the class of interest, and a large quantity of unlabeled and unknown data. A classification boundary differentiating the underlying positive and negative classes is still desired. This is known as the positive unlabeled learning problem (PU learning). An illustration comparing standard classification and the PU learning problem is shown in Figure 1.

Many problems fit into this framework. Medical applications including identifying or priority ranking genes or gene combinations that influence disease incidence. Some of the earlier uses for PU learning included text, email, and web-page classification. Security and signal processing applications are severely underrepresented in the Positive and Unlabeled learning domain. Initial forays into image classification have explored satellite image land-type classification and facial authentication. Classifying satellite images, radar images, and others is a natural fit for PU learning. See [3] for more examples and for references to the above-mentioned examples.

In this project, we first performed a literature review of the existing work in this area and examples of its use and possible future applications [3]. We then introduced a new probabilistic algorithm that uses a modified logistic regression (MLR) to solve the PU learning problem with a high degree of accuracy [4]. This work is an extension of We are currently in the process of designing a further improved PU learning algorithm using ensemble techniques called pBMLR [5].

In testing our algorithms, we have compared them against some of the current state of the art [6], [7] over both simulated datasets and real-world image datasets including the MNIST dataset. Our results have shown that our approach and algorithms produce substantially improved results over existing solutions. [4], [5]



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

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

Figure 1: Illustration comparing PU learning and traditional binary classification problem

[1] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2011.

[2] U. S. Shanthamallu, A. Spanias, C. Tepedelenlioglu, and M. Stanley, "A brief survey of machine learning methods and their sensor and IoT applications," in *IEEE IISA*, Larnaca, Cyprus, Aug. 2017.

[3] K. Jaskie and A. S. Spanias, "Positive and Unlabeled Learning Algorithms and Applications : A Survey," in *IEEE IISA*, Patras, Greece, IEEE, Jul. 2019.

[4] 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.

[5] K. Jaskie and A. Spanias, "pBMLR: A Pre-Boosted Modified Logistic Regression for Positive Unlabeled Learning." (In Progress)

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

[7] M. C. Du Plessis and M. Sugiyama, "Class Prior Estimation from Positive and Unlabeled Data," *IEICE Trans. Inf. Syst.*, vol. E96-D, no. 5, pp. 1358–1362, 2014.