A MODIFIED LOGISTIC REGRESSION FOR POSITIVE AND UNLABELED LEARNING

Kristen Jaskie[1], *Charles Elkan*[2], *Andreas Spanias*[1]

[1]SenSIP Center, School of ECEE, Arizona State University, USA

[2]Computer Science and Engineering, University of California San Diego, USA

ABSTRACT

The positive and unlabeled learning problem is a semi-supervised binary classification problem. In PU learning, only an unknown percentage of positive samples are known, while the remaining samples, both positive and negative, are unknown. We wish to learn a decision boundary that separates the positive and negative data distributions. In this paper, we build on an existing popular probabilistic positive unlabeled learning algorithm and introduce a new modified logistic regression learner with a variable upper bound that we argue provides a better theoretical solution for this problem. We then apply this solution to both simulated data and to a simple image classification problem using the MNIST dataset with significantly improved results.

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

1. INTRODUCTION

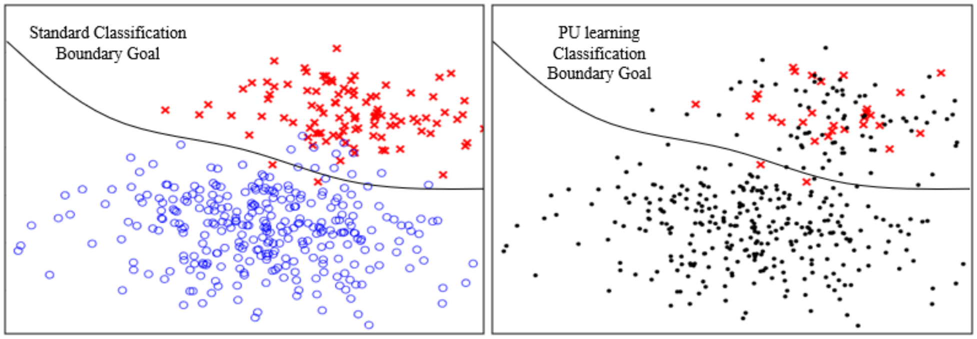
Classification is an important task in machine learning and signal processing. Image object identification, video activity categorization, and acoustic classification of audio signals are examples of signal processing classification problems. Text and document classification, fraud detection, and disease gene identification are other traditional machine learning classification problems.

In the standard supervised classification problem, we are given a large quantity of training data samples x, each labeled with its associated binary class y - typically positive ($y = 1$) or negative ($y = 0$). A classification algorithm learns a model $f(x)$ from the features of these labeled training samples. Given a new sample with no label, this model then classifies that unlabeled sample as belonging to either the positive or negative class. Several algorithms are used for this problem, including logistic regression, support vector machines (SVMs), and Artificial Neural Networks (ANNs).

In the real world, it is often quite difficult and/or expensive to gather sufficient quantities of labeled data for training. It has been shown that using additional cheap, unlabeled data in training classifiers can help solve this problem [1]. Partially supervised learning algorithms can make use of small sets of positive and negative labeled training data and large sets of unlabeled training data. Frequently, however, even a small set of negative training data is unavailable and learning from only positive and unlabeled samples becomes desirable. This is known as the Positive and Unlabeled learning problem (PU Learning). A simple illustration is given in Figure 1. Notice that only a small proportion of data samples are labeled.

One application of the positive and unlabeled learning problem is image object classification and detection. Image object classification can be used to detect man-made structures in satellite images (such as sites of archeological or military interest), identify a particular make and model of vehicle in images of cars in a city or on a freeway, or to classify an image from a dashcam of an autonomous vehicle as something requiring an emergency stop such as a person, object, or large animal in the road. Additional applications are described in [2].

In this paper, we propose a probabilistic algorithm that uses a modified logistic regression to solve the positive unlabeled learning problem. This approach involves a modification to the benchmark algorithm for this problem proposed by Elkan and Noto in [3]. Additional solutions to this problem are discussed in [2] and [4]. We discuss the theoretical justification for the modification and apply the algorithm to simulated data and images from the MNIST image dataset.



(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.

Figure 1: Illustration of the Positive and Unlabeled binary classification problem.

2. ASSUMPTIONS

Learning the overall percentage of positive samples $p(y = 1)$, also known as the class prior, directly from positive and unlabeled data is an important problem, first identified in [5] and [6]. Without expert domain knowledge, it is impossible to know if we have well separated homogeneous positive and negative classes with a low probability that a positive sample is labeled, or poorly separated positive and negative classes with significant overlap and a high probability that a positive sample is labeled. Both scenarios can result in the same positive and unlabeled set. This is illustrated in Figure 2. Notice that in both cases, the labeled and unlabeled data samples are completely overlapping and non-separable as illustrated in Figure 1b. This leads us to two important and necessary assumptions.

First, we make the implicit and necessary assumption that the region of highest density of labeled positive samples must consist entirely of positive samples—i.e. that it is homogeneous as shown in Figure 2. Where the density of known positive samples decreases, we conclude the overlap of positive and negative distributions. This is sometimes called the partial separability or positive subdomain assumption [4].

Second, we explicitly assume that our positive labeled samples are "selected completely at random" from the set of all positive samples. This is the SCAR assumption, first described in [3], and it is necessary if we are to draw reasonably accurate conclusions about the positive and negative distributions of our dataset. To see this, let us assume that our positively labeled samples are NOT selected completely at random from the set of all positive samples. That is, suppose there is some bias in the sampling process. This bias is then going to be learned from the training set and passed along to our model.

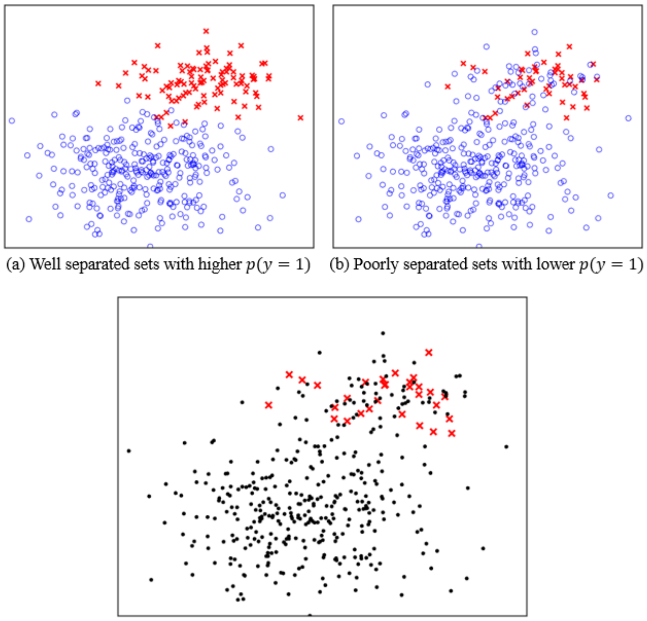
3. PROPOSED METHOD – THE THEORY

In this section, we introduce the mathematics behind our solution. Notice that because of the SCAR assumption described in section 2, the probability that a data sample is labeled positive given it’s feature characteristics, will never be 1, but will instead be some unknown constant c. This means that we need a probabilistic learner that has a maximum value of c instead of 1. Our solution is to construct and use a Modified Logistic Regression. The following sections describe this in more detail.

3.1 A Probabilistic Approach

A probabilistic solution for estimating $p(y = 1|x)$ and with it, $p(y = 1)$ is derived in [3]. First, a new random variable s is introduced that represents whether a sample is labeled or unlabeled. If a sample is labeled positive, then $s = 1$. If the sample is unlabeled, meaning it is unknown whether it is positive or negative, then $s = 0$. The positive and unlabeled problem can be stated formally using this notation as

$$p(s = 1 \mid y = 0) = 0.$$



(a) Well separated sets with higher $p(y = 1)$ (b) Poorly separated sets with lower $p(y = 1)$

(c) Both of the above scenarios (a) and (b) could produce a positive and unlabeled set like this.

Figure 2: Separability Assumptions.

Using this extra variable s, the assumption that the positive labeled data are selected completely at random from the set of all positive data, i.e. the SCAR assumption, can be stated formally as

$$p(s = 1 \mid x, y = 1) = p(s = 1 \mid y = 1) = c. \qquad (1)$$

Here, $c = p(s = 1|y = 1)$ is the constant probability that a positive sample is labeled.

The following derivation gives us the relationship between $p(y = 1|x)$, $p(s = 1|x)$ and c and is found in [3]. Given equation (1), the likelihood, or conditional probability, that a sample x belongs to the positive set $p(y = 1|x)$ can be derived as follows.

$$\begin{aligned} p(s = 1 \mid x) &= p(s = 1 \wedge y = 1 \mid x) \\ &= p(y = 1 \mid x)\, p(s = 1|y = 1, x) \\ &= p(y = 1|x)\, p(s = 1|y = 1) \\ &= p(y = 1|x)\, c. \end{aligned}$$

Therefore,

$$p(y = 1|x) = p(s = 1|x)/c. \qquad (2)$$

Here, $f(x) = p(y = 1|x)$ is known as a “traditional classifier” and $g(x) = p(s = 1|x)$ is called a “nontraditional classifier” in that it learns the label of a sample rather than learning if it is positive. There are several important consequences of this derivation. As [3] point out, $f(x)$ is an increasing function of $g(x)$. Additionaly, $f(x) = g(x)/c$ is a well-defined probability, such that $f(x) \leq 1$ only if $g(x) \leq c = p(s = 1|y = 1)$. Formally,

$$g(x) = p(s = 1 \mid x) \leq p(s = 1|y = 1) = c. \qquad (3)$$

Unfortunately, learning $g(x) = p(s = 1|x)$ is inherently difficult due to the inseparability of the labeled and unlabeled classes due the SCAR assumption. [3] used standard logistic regression to learn the non-traditional classifier $g(x)$. In this paper, we argue that a modified logistic regression will provide a better classifier for $g(x)$. In the next section, we describe this modified logistic regression in detail.

3.2 Modified Logistic Regression

In Section 31, we described a derivation of $p(y = 1|x)$. If we could learn an estimate of c and the nontraditional classifier $g(x) = p(s = 1|x)$, we could calculate $f(x) = p(y = 1|x)$ from equation (2). As mentioned above, learning the nontraditional classifier $g(x)$ is inherently difficult due to the SCAR assumption and the overlapping nature of our labeled and unlabeled classes.

As shown in equation (3), for $f(x) = p(y = 1|x)$ to be a well-defined probability, $g(x) = p(s = 1|\, x) \leq p(s = 1|\, y = 1) = c$. A critical point therefore, is that the probability output for our nontraditional classifier $g(x) = p(s = 1|x)$ needs to be in the range of $[0, c]$ not $[0,1]$ as are typical probabilities generated by standard probabilistic learning algorithms. Our solution is to introduce a modified logistic regression algorithm with a variable upper bound that is less than or equal to 1. This is a simplification of a Generalized Logistics Curve, also known as a Richard's Curve. The upper bound is not given as an input to the algorithm but is learned as part of the training process. From (3) we know that if $g(x) = p(s = 1|x)$ is well calibrated, it will be less than or equal to c. It follows that we can take the asymptote of our modified logistic regression to be an estimate of c, $\hat{c}$.

$$\hat{c} = \max_{x} p(s = 1|x)$$

A standard probabilistic logistic regression learner of $g(x)$ that asymptotes at 0 and 1 has the equation:

$$g_{SLR}(x) = \; p(s = 1\,|\bar{x}\,) = \frac{1}{1 + e^{-\bar{w}\cdot\bar{x}}}$$

where $\bar{w}$ is the learned d-dimensional weight vector and $\bar{x}$ and s are the input feature values and binary label for the given sample, respectively. To enable an upper bound of less than 1, we insert an additional variable in the denominator and differentiate our learner by naming it $g_{MLR}(x)$. We can ensure that this upper bound is greater than or equal to zero by squaring this variable: b^2. Because b^2 is in the denominator, the upper bound will be less than or equal to one.

$$g_{MLR}(\bar{x}) = p(s = 1|\bar{x}) = \frac{1}{1+b^2+e^{-\bar{w}\,\bar{x}}} \tag{4}$$

The upper bound asymptote can be calculated as

$$\hat{c} = 1/(1 + b^2) \tag{5}$$

An illustration of $\mathrm{g}_{\mathrm{MLR}}(x)$ with an asymptote at a c value less than 1 can be seen in Figure 3. This modified logistic regression leads us to our algorithm, described in the next section.

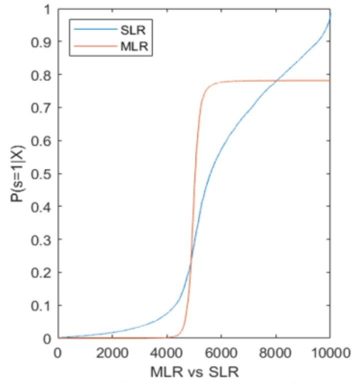


Figure 3: Modified logistic regression (MLR) compared to standard (SLR) when learning $g(x) = p(s = 1|x)$ from equation (4) on simulated data where $c = 0.75$.

4. PROPOSED METHOD – THE ALGORITHM

The goal of any binary classifier is to assign each data sample into one of two classes, depending on which is more likely given the features available. In a probabilistic algorithm, we go one step further by returning the probability that each data sample belongs in the positive class. In this algorithm, we use the modified logistic regression algorithm described in the previous section.

4.1. Step 1 – Learning a nontraditional classifier

The first step in our algorithm will be to learn a non-traditional classifier from our training data to identify the probability that a given sample x is labeled positive (not that it IS positive). That is, we would like to learn $p(s = 1|x)$ and $p(s = 0|x)$. From equation (4) above, we get that

$$p(s = 1\,|\bar{x}\,) = \frac{1}{1 + b^2 + e^{-\bar{w}\cdot\bar{x}}}$$

From this, we can construct $p(s = 0|x)$ as

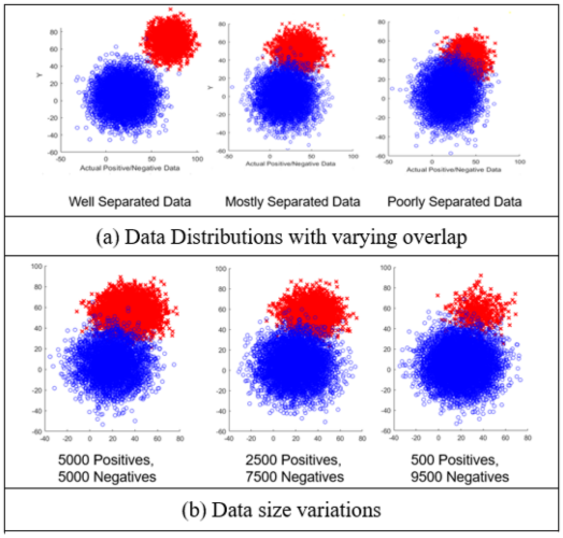
$$p(s = 0|\,\bar{x}\,) = 1 - p(s = 1|\bar{x}) = \frac{b^2 + e^{-\bar{w}\cdot\bar{x}}}{1 + b^2 + e^{-\bar{w}\cdot\bar{x}}}.$$

Our first step is to learn the values of the weight vectors $\bar{w}$ and the random variable b that maximize the likelihood of each data sample and label. To maximize this likelihood during training, we take the gradient of the log-likelihood and train for ε epochs with an adaptive learning rate λ until convergence. The values for ε and λ will need to be tuned individually for each data set though $\epsilon = 1000$ and $\lambda = 1/\epsilon$ seem to be fairly effective in most cases.

4.2. Step 2 – Construct the final classifier

Once b has been learned in the previous step, we can use equation (5) to estimate c as $\hat{c} = 1/(1 + b^2)$. With an estimate of c, we can now estimate our end-goal, the traditional classifier $p(y = 1|x)$ using $p(s = 1|x)$ and the estimate of c, $\hat{c}$ as

$$p(y = 1|x) = \frac{p(s = 1|x)}{\hat{c}}.$$



(a) Data Distributions with varying overlap

(b) Data size variations

Figure 4: Simulated Data Distribution.

5. EXPERIMENTAL DATA AND RESULTS

To test the proposed modified logistic regression algorithm, we used both simulated data as well as some basic image classification using the MNIST dataset. Due to the potential for uneven class sizes, the F-score was used as our evaluation metric as accuracy and error rate metrics are not useful when class sizes are heavily skewed.

5.1 Simulated Data Setup and Results

To determine the effectiveness of our modified logistic regression algorithm, we created three simulated data distributions in two dimensions shown in Figure 4(a). For each of these distributions, we used three different data sizes shown in Figure 4(b). Finally, with each of these combinations of data distribution and data size, we looked at nine different values of c from 0.1 to 0.9. Recall that c is the percentage of known positives out of the total positives. This gives us 81 different scenarios on which to evaluate our solution. For each test scenario, we performed 50 Monte Carlo simulations and recorded their average.

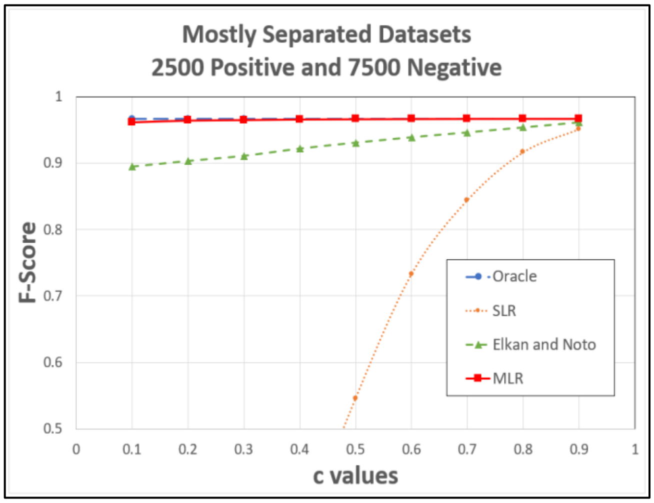


Figure 5: Results over selected simulated dataset with uneven class sizes.

In performing these simulations, we compared an Oracle (classification using a standard logistic regression where *all the true labels are known*), standard logistic regression (SLR) on the PU data, the three estimators presented in [3], and our proposed modified logistic regression (MLR) algorithm. For illustrative simplicity, we have only plotted the best estimator from [3] which we have called 'Elkan and Noto' after the authors.

We found that our proposed MLR algorithm gave improved results 96.3% of the time, or in 78 out of 81 simulations.

5.2 MNIST Data and Results

We also compared the algorithms described in 5.1 to the image processing MNIST problem of handwritten digit classification. For simplicity, we performed no feature engineering, and simply used the unrolled pixel values given for all algorithms. This standard classification problem has 60,000 training samples of the digits 0-9 with approximately 6000 samples for each digit. We performed binary classification of the most commonly mis-classified digit pairs: 3 and 5, 3 and 8, and 5 and 8. We tested c values of 0.1, 0.25, 0.5, and 0.75.

In *every case*, our proposed MLR outperformed the other algorithms tested, by an average of over 17%.

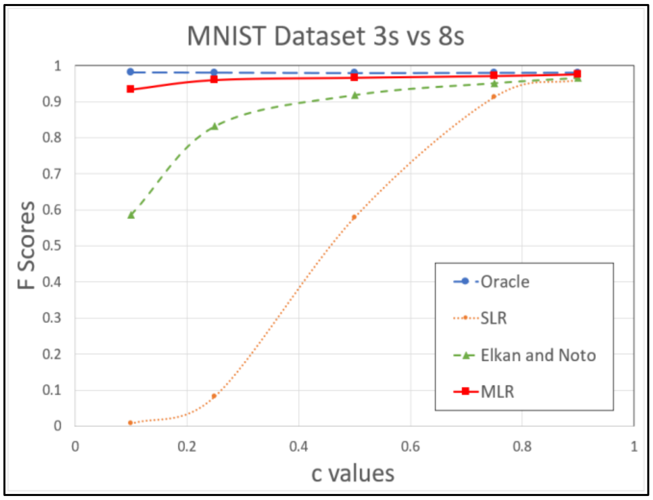


Figure 6: MNIST Results between easily confusable digits 3 and 8. When $c = 0.1$, only 10% of the 3's are known and all others (90% of 3's and 100% of 8's) are unknown.

6. CONCLUSION

The positive unlabeled (PU) learning problem exists in many important real-world applications. Labeled data is difficult and expensive to obtain, and true negatively labeled data may be impossible or impractical. Our solution introduces a modified logistic regression and has been shown to be extremely effective and to out-perform the current state-of-the art algorithms on both simulated data and when using the MNIST dataset for real-world image classification. Future work will compare more algorithms over additional datasets.

7. ADDITIONAL RESOURCES

The positive unlabeled learning problem is a growing area of research in semi-supervised learning. Recent survey papers [2], [4] give an overview of the field, its applications, and approaches. Some recent papers of interest include [7]–[22], and [23]–[25] are foundational. Previous work in the SenSIP Center in machine learning methods includes [26]–[29]

8. AKNOWLEDGEMENT

This project was funded in part by the ASU SenSIP Center and the NSF I/UCRC award 1540040.

9. REFERENCES

[1] F. De Comité, F. Denis, R. Gilleron, et all, "Positive and unlabeled examples help learning," *Algorithic Learn. Theory*, v. 1720, pp. 219–230, Dec. 1999.

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

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

[4] J. Bekker and J. Davis, "Learning From Positive and Unlabeled Data: A Survey," *ArXiv*, Nov. 2018.

[5] F. Denis, R. Gilleron, and F. Letouzey, "Learning from positive and unlabeled examples," *Theor. Comput. Sci.*, vol. 348, no. 1, pp. 70–83, Dec. 2005.

[6] D. Zhang and W. S. Lee, "A Simple Probabilistic Approach to Learning from Positive and Unlabeled Examples," in *UKCI*, Springer, pp. 83–87, Sep. 2005.

[7] M. Kato, T. Teshima, and J. Honda, "Learning from Positive and Unlabeled Data with a Selection Bias," *ICLR*, May 2019.

[8] J. Zhang, Z. Wang, J. Meng, Y. P. Tan, and J. Yuan, "Boosting positive and unlabeled learning for anomaly detection with multi-features," *IEEE Trans. Multimed.*, vol. 21, no. 5, pp. 1332–1344, May 2019.

[9] G. Niu, M. C. du Plessis, T. Sakai, Y. Ma, and M. Sugiyama, "Theoretical Comparisons of Positive-Unlabeled Learning against Positive-Negative Learning," in *NIPS*, Barcelona, Spain, ACM, pp. 1207–15, Dec. 2016.

[10] M. Claesen, F. De Smet, J. A. K. Suykens, and B. De Moor, "A robust ensemble approach to learn from positive and unlabeled data using SVM base models," *Neurocomputing*, vol. 160, pp. 73–84, Jul. 2015.

[11] M. C. Du Plessis, G. Niu, and M. Sugiyama, "Class-prior estimation for learning from positive and unlabeled data," in *Asian Conf. Mach. Learn.*, Hong Kong, China, pp. 221–236, Nov. 2015.

[12] M. C. du Plessis, G. Niu, and M. Sugiyama, "Convex Formulation for Learning from Positive and Unlabeled Data," in *ACML*, Hong Kong, China, Springer, pp. 221–236, Nov. 2015.

[13] M. C. du Plessis, G. Niu, and M. Sugiyama, "Analysis of Learning from Positive and Unlabeled Data," *NIPS*, pp. 703–711, Dec. 2014.

[14] F. Mordelet and J. P. Vert, "A bagging SVM to learn from positive and unlabeled examples," *Pattern Recognit. Lett.*, vol. 36, pp. 201–209, 2014.

[15] J. Bekker and J. Davis, "Beyond the Selected Completely At Random Assumption for Learning from Positive and Unlabeled Data," *ArXiv*, Jun. 2018.

[16] L. de Carvalho Pagliosa and R. F. de Mello, "Semi-supervised time series classification on positive and unlabeled problems using cross-recurrence quantification analysis," *Pattern Recognit.*, vol. 80, pp. 53–63, Aug. 2018.

[17] M. Hou, B. Chaib-Draa, C. Li, and Q. Zhao, "Generative Adversarial Positive-Unlabelled Learning," in *IJCAI*, Stockholm, Sweden, Jul. 2018.

[18] E. Sansone, F. G. B. De Natale, and Z. H. Zhou, "Efficient Training for Positive Unlabeled Learning," *TPAMI*, Jul. 2018.

[19] H. Gan, Y. Zhang, and Q. Song, "Bayesian belief network for positive unlabeled learning with uncertainty," *Pattern Recognit. Lett.*, Apr. 2017.

[20] R. Kiryo, G. Niu, M. C. du Plessis, and M. Sugiyama, "Positive-Unlabeled Learning with Non-Negative Risk Estimator," in *NIPS*, Long Beach, Curran Assoc. Inc., pp. 1674–84, Dec. 2017.

[21] J. Zhang, Z. Wang, J. Yuan, and Y.-P. Tan, "Positive and Unlabeled Learning for Anomaly Detection with Multi-features," in *ACM MM*, Mountain View, ACM, pp. 854–62, Oct. 2017.

[22] D. Ienco and R. G. Pensa, "Positive and unlabeled learning in categorical data," *Neurocomputing*, vol. 196, pp. 113–124, Jul. 2016.

[23] G. Ward, T. Hastie, S. Barry, J. Elith, et all, "Presence-only data and the em algorithm," *Biometrics*, vol. 65, no. 2, pp. 554–63, May 2009.

[24] B. Z. Zhang and W. L. Zuo, "Co-EM support vector machine based text classification from positive and unlabeled examples," in *ICINIS*, Wuhan, China, IEEE, pp. 745–748, Nov. 2008.

[25] F. Denis, R. Gilleron, and M. Tommasi, "Text classification from positive and unlabeled examples," in *IPMU*, Annecy, France, Jul. 2002.

[26] 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.

[27] U. S. Shanthamallu, J. J. Thiagarajan, H. Song, and A. Spanias, "GrAMME: Semi-Supervised Learning using Multi-layered Graph Attention Models," *IEEE Trans. Neural Networks Learn. Syst.*, pp. 1–12, Nov. 2019.

[28] H. Song, J. J. Thiagarajan, P. Sattigeri, and A. Spanias, "Optimizing kernel machines using deep learning," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 11, pp. 5528–5540, Feb. 2018.

[29] A. Wisler, V. Berisha, A. Spanias, and A. O. Hero, "Direct estimation of density functionals using a polynomial basis," *IEEE Trans. Signal Process.*, vol. 66, no. 3, pp. 558–572, Feb. 2018.