BMC Bioinformatics

RESEARCH

Open Access

Leveraging permutation testing to assess confidence in positive-unlabeled learning applied to high-dimensional biological datasets

Shiwei Xu[1] and Margaret E. Ackerman[1,2,3*]

*Correspondence: margaret.e.ackerman@dartmouth.edu

[3] Thayer School of Engineering, Dartmouth College, 14 Engineering Dr., Hanover, NH 03755, USA
Full list of author information is available at the end of the article

Abstract

Background: Compared to traditional supervised machine learning approaches employing fully labeled samples, positive-unlabeled (PU) learning techniques aim to classify "unlabeled" samples based on a smaller proportion of known positive examples. This more challenging modeling goal reflects many real-world scenarios in which negative examples are not available—posing direct challenges to defining prediction accuracy and robustness. While several studies have evaluated predictions learned from only definitive positive examples, few have investigated whether correct classification of a high proportion of known positives (KP) samples from among unlabeled samples can act as a surrogate to indicate model quality.

Results: In this study, we report a novel methodology combining multiple established PU learning-based strategies with permutation testing to evaluate the potential of KP samples to accurately classify unlabeled samples without using "ground truth" positive and negative labels for validation. Multivariate synthetic and real-world high-dimensional benchmark datasets were employed to demonstrate the suitability of the proposed pipeline to provide evidence of model robustness across varied underlying ground truth class label compositions among the unlabeled set and with different proportions of KP examples. Comparisons between model performance with actual and permuted labels could be used to distinguish reliable from unreliable models.

Conclusions: As in fully supervised machine learning, permutation testing offers a means to set a baseline "no-information rate" benchmark in the context of semi-supervised PU learning inference tasks—providing a standard against which model performance can be compared.

Keywords: Positive-unlabeled learning, Semi-supervised machine learning, High-dimensional biological data, Permutation testing

Introduction

Classification and clustering with machine learning algorithms are approaches that have been widely applied in biological and biomedical research. However, given significant experimental advances, these datasets now often present the "curse of dimensionality" [1]. While thousands of features can be easily collected in profiling studies, the number

of samples is generally much smaller and often constrained by resources or biological rarity. Particularly in the context of such "wide" datasets, care must be taken to rigorously moderate overfitting by developing effective methods to reduce the impact of redundant and non-informative features, such as by optimizing model building in the context of cross-validation and regularization. Permutation testing, in which the label or value to be modeled is scrambled among samples prior to modeling, presents an additional means to assess model robustness [2]. Permutation and modeling steps can then be repeated to generate a distribution that reflects the probability of achieving of model of a given quality by chance [3]. This distribution can be used to set a benchmark against which performance of the model trained on actual labels or values can be evaluated.

Here, we explore the use of permutation testing to address the challenging problem of establishing confidence in modeling results in the setting of semi-supervised learning (SSL) classification tasks. Specifically, we investigate a subclass of SSL tasks, known as positive unlabeled (PU) Learning, which has attracted the attention of researchers with datasets that consist of a small proportion of labeled positive examples and a vast majority of unlabeled samples that contain both positive and negative samples, but among which no definitive examples of true negatives are known. Summarized by Li et al. [4], major PU learning algorithms developed for bioinformatics tasks can be categorized into Reliable Negative Selection and Base Classifier adaptation, including but not limited to PU strategies such as bootstrapped aggregation (Bagging). Previously, we studied the empirical behavior of the transductive PU Bagging algorithm in high dimensional datasets with varied group separation and label imbalance [5]. In this context, prediction performance for PU bagging was superior to both a single biased SVM classifier, which is a frequently used method in PU learning tasks, especially when the proportion of known positive (KP) samples among all samples is small. PU bagging also outperformed traditional fully supervised (two-class) models in which unlabeled samples were modeled as the negative class but performance was evaluated against ground truth class labels [5]. Furthermore, by comparing multiple types of machine learning classifiers in the PU bagging approach, we showed that the algorithm was relatively insensitive to the choice of classifier, demonstrating an advantage when there is no ground truth label available to contribute to optimizing model selection.

Most prior work comparing multiple PU learning-based approaches reported binary prediction performance with evaluation metrics such as accuracy, F1 score, and Matthew's correlation coefficient (MCC), each of which requires ground truth reference labels or "True Negative" examples for validation. However, validating prediction outcomes is one of the major barriers in PU learning applications for real-world use cases, especially in biological and biomedical tasks for which identification of "true negative" examples is impractical, unethical, or even impossible. Likewise, most state-of-the-art PU approaches, the ensembled model will classify the samples whether or not "true negatives", or separable underlying clusters are present. This challenge to validation is considered to be one of the major factors that limits confidence in PU learning results. In practice, there have been a limited number of studies in which researchers have attempted to validate predictions based on positive examples only (with no definitive negative examples) [6, 7]. A metric termed explicit precision recall (EPR) scored prediction quality by calculating the proportion of known positive samples that were

predicted positive [7]. However, whether "good" prediction of positive examples can act as a reliable surrogate for "good" prediction of negative samples is certain to be context-dependent, and this score lacks a means to disincentivize calling all samples positive, which will yield the greatest EPR score, but not a reliable model in cases where there are two classes.

To this end, we propose a generalizable methodology to evaluate the robustness of positive examples to classify unlabeled samples in the absence of negative examples in the setting of transductive PU learning by integrating a previously developed PU strategy termed "Spy Positive Technique" with the PU Bagging algorithm. Proposed by Liu et al., the spy positive technique was developed to determine a confident decision boundary in PU methods that are categorized as "Reliable Negative Selection" approaches [4, 8]. In this method, a small proportion of known positive (KP) examples are randomly sampled into the unlabeled group and their probability to be classified as positive or negative class evaluated. Although Bekker and Davis [9] argued that the method might not be adequate to determine decision boundaries for datasets without a "sufficient" number of labeled samples, we found it suitable to evaluate the entire known positive (KP) set in PU bagging-based classification by repeatedly treating a small portion of KP as unlabeled and scoring them in the bagging procedure with other U set samples. Moreover, we explore the use of combining this positive set evaluation approach and permutation tests to address the challenging problem of establishing confidence in modeling results in the setting of SSL classification tasks relevant to biological and biomedical research. Overall, while the new approach we report cannot definitively establish the reliability of modeling results, it can be used to define the probability that similar classification performance could be achieved by chance, giving a means whereby the results of PU learning predictions can be rejected or supported for further biological investigation.

Results

Application of PU learning to datasets with and without underlying class differences

Among semi-supervised machine learning approaches, which aim to classify positive and negative examples with incomplete label information, PU learning is a set of methods that seek to classify samples when known examples of only one class exists, typically among a larger set of unlabeled samples. These methods provide value across a wide range of fields (Fig. 1A), varying from spam or fake comment identification for filtering [10, 11] and search result optimization using "likes" or viewing time information [12], among others. Biological applications include virtual drug and gene screening [13], species presence prediction for ecological monitoring [14], embryo selection for assisted reproduction [15], and protein–protein interaction prediction [16]. We have demonstrated the ability of PU learning to contribute to predicting protection status in vaccine efficacy field trials [17], in which positive and negative classes in this application reflect unprotected and protected classes, respectively. In this use case, definitive class labels are only available for a subset of the positive (known P class) subjects, those who were exposed and infected; whereas the uninfected class is comprised of both protected and unprotected subjects who were simply not exposed to the pathogen, are unlabeled (U class). This approach has demonstrated the ability to discover CoP missed by conventional positive/negative (infected/uninfected) class analysis [17]. Importantly, however,

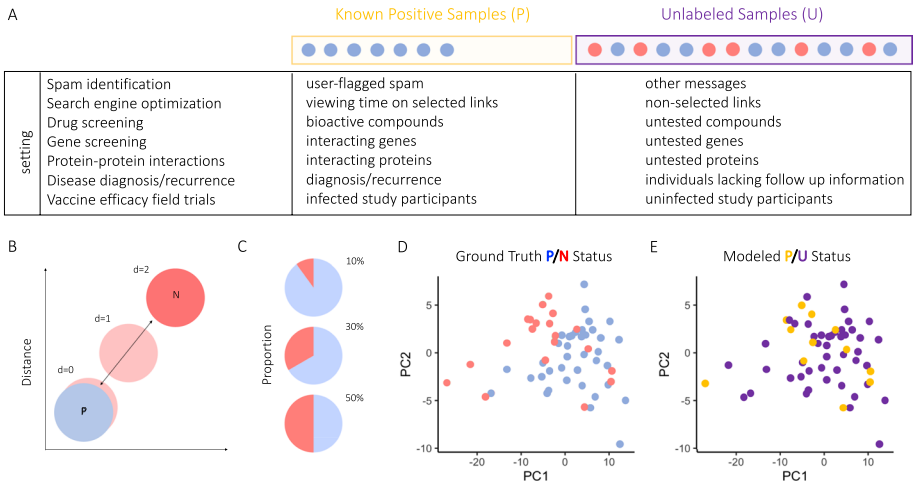

Fig. 1 PU learning use cases, synthetic data generation, data unlabeling. **A** Examples of settings in which PU learning may improve class prediction. **B** Synthetic datasets in which there is no (distance d = 0), low (d = 1), or high (d = 2) degree of difference between positive (P) and negative (N) classes were generated. **C** The unlabeled (U) sample set was comprised of varying proportions (10%, 30%, 50%) of N among U samples. **D** Exemplary visualization of P and N sample data profiles. **E** Exemplary partial relabeling of P/N class data to generate P and U classes

real world applications lack ground truth protection status information to validate models, and would benefit from an analytical control.

Here, we generated nine high dimensional (features, $p \geq n$ samples) synthetic datasets with different hypercube distances (high (class separation = 2), medium (class separation = 1) and a negative control (class separation = 0) (Fig. 1B) with varied ground truth label class ratios (10%, 30%, 50% True Negative) (Fig. 1C). Lastly, the Wisconsin Breast Cancer Database [18], The Cancer Genome Atlas [19], and a nonhuman primate vaccine study datasets [20], each known to have at least some real biological signal between P and N classes were evaluated (ie: Fig. 1D). Each dataset was partially unlabeled to present P and U classes of differing proportions (ie: Fig. 1E). High, low, and no class separation datasets were then used to evaluate metrics that could define model confidence. By comparing results between actual and permuted class labels, and between input data sets that were or were not comprised of two actually distinct classes, we aimed to calculate the likelihood that a given model performance would be observed at random.

Generating a distribution of positive class probability scores that reflects the null hypothesis

Inspired by the "spy" positive sample technique that was previously developed in two-step reliable negative approaches (Fig. 2A), we score the class 1 probability of each "spy fold" with the PU Bagging algorithm (Fig. 2B) [8]. The class 1 probability of each positive sample when modeled as a "spy" is further aggregated by two different scoring methods and pooled under a repeated cross-validated spy fold strategy [21]. The resulting distribution of PU Bagging scores for actual positive samples can then be compared to the distributions observed when the same inference strategy is applied after label permutation (Fig. 2B).

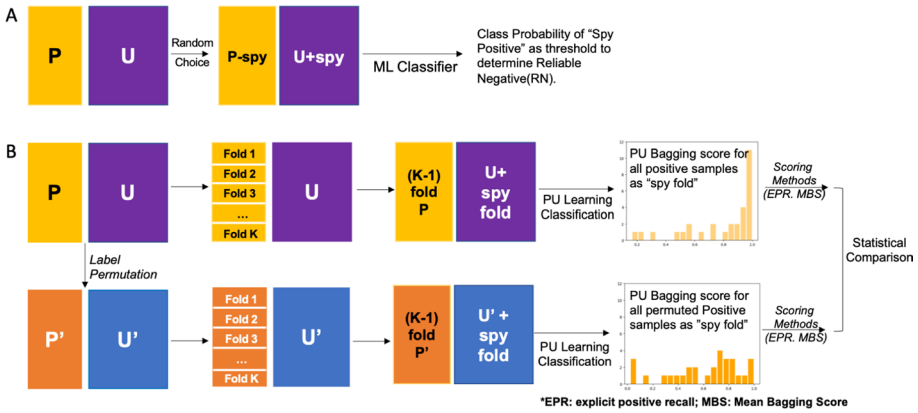

Fig. 2 Graphical representation of the application of permutation and "spy" techniques in PU learning. **A** Proposed by Liu et al. [8], the spy method randomly samples a small percentage of the definitively labeled positive (P) class and mixes them into the unlabeled (U) set as "spies". A classifier is then trained on the basis of the remainder of the P samples and expanded U set. The class probability of the "spy" samples is then employed to set the threshold for identification of the reliable negative (RN) set in a *Two-Step Reliable Negative* PU strategy. **B** Implementation of permutation testing to evaluate confidence in PU learning classification. PU bagging scores across replicates and folds are calculated for "spies" (top) and compared to scores observed for "spies" when P and U labels were randomly permuted (bottom) using explicit positive recall (EPR) [7], and mean bagging scores (MBS)

Evaluating confidence in class label inferences based on synthetic data

In previous work, we identified two characteristics that generally negatively impacted the prediction performance of PU learning methods: small separation between samples associated with ground truth label classes, and a low number of known positives among all underlying positive sample [17]. Here, we employed nine high dimensional ($p \geq n$) synthetic datasets with different hypercube distances (high (class separation = 2), low (class separation = 1), and no separation (class separation = 0). These classes were combined with varied ground truth label class ratios (10%, 30%, 50% True Negative) to form the unlabeled set. The varying difficulty of these tasks, based on the degree of class distinction in the underlying data captured in the first two principal components (PCs), are depicted for ground truth and positive-unlabeled classes for varying degrees of class separation, with the no class separation condition serving as a negative control (Fig. 3A).

We calculated area under the receiver operator characteristic curve (ROC-AUC) between the modeled class 1 probability for known positive examples and the ground truth label for the unlabeled set in order to define the expected prediction performance of inferring class assignments for the unlabeled set (Fig. 3B inset). As expected, the unlabeled set AUC (U-AUC) values varied from excellent (1.00) to close to nominally random performance (0.64) in association with the degree of class separation when a high proportion (50%) of true negative samples were included in the unlabeled set. When the proportion of true negative samples within the unlabeled set was decreased to 10%, the U-AUC values for each condition showed a consistent reduction and exhibited the expected variation based on the level of class separation, ranging from 0.88 (excellent) to 0.54 (no discrimination) [22].

Modeling was then repeated after the actual PU labels were permuted multiple times. The PU bagging scores for actual and permuted known positive samples were then calculated and compared to define confidence in modeling the definitively labeled class

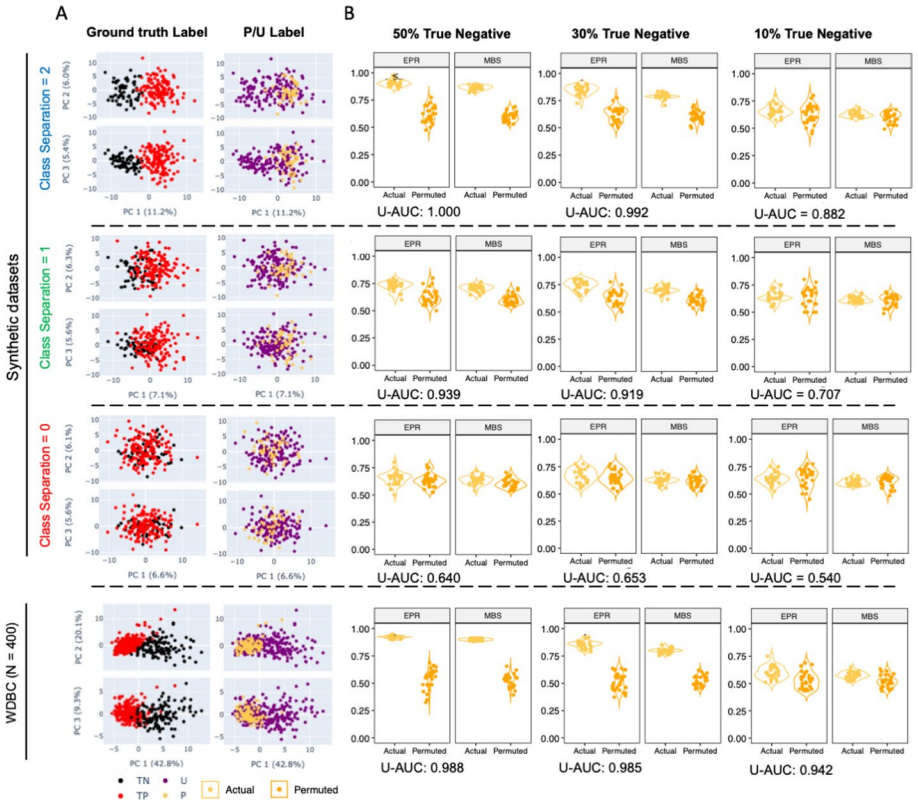

Fig. 3 Permutation testing to evaluate model robustness across varying class separations and with extreme class imbalance. **A** Data visualization (30% True Negative) using first three components from PCA, labeled by both ground truth class labels (left, black and red) and P/U labels (right, purple and yellow) in the simulation settings for synthetic data with 3 different degrees of class separation (top) and Wisconsin Diagnostic Breast Cancer (WDBC) dataset (bottom). **B** Violin plots comparing the Positive Set score under actual and permuted P/U labels for varying portions (50–10%, left to right) of True Negatives among the U set. Explicit positive recall (EPR) and mean bagging scores (MBS) are presented from models resulting for actual and permuted labels. Area under the curve for classification of unlabeled samples (U-AUC) is reported in below each test condition

samples (Fig. 3B). To thoroughly evaluate performance for the samples labeled as positive, both explicit positive recall (EPR) [7] and mean bagging scores (MBS) were calculated. In these simulations, we hypothesized that if the actual positive samples exhibited a distinct profile, the distribution of scores learned for actual known positives would differ from those learned from modeling with a number-matched set of samples drawn at random (Fig. 4A). Importantly, for the dataset with no class separation, the distribution of z-score p values was similar to the null hypothesis distribution (Supplemental Fig. S1), and in 1000 replicates, fewer than 1% of calculated p values fell below the nominal threshold of $p < 0.05$. In contrast, when classes were well separated (class separation = 2), distributions for actual and permuted positive sample EPR and MBS scores were distinct by Cliff's Delta estimate, which is a measure of effect size (Fig. 4B), as well as for p values derived from z-scores (Fig. 4C), as long as sufficient true negatives were present among the unlabeled set (Fig. 4). As expected, performance was superior for the easier inference tasks, with high class separation and a high number (and proportion) of true negatives. In scenarios with a low proportion (10%) of true negatives, EPR and MBS values for

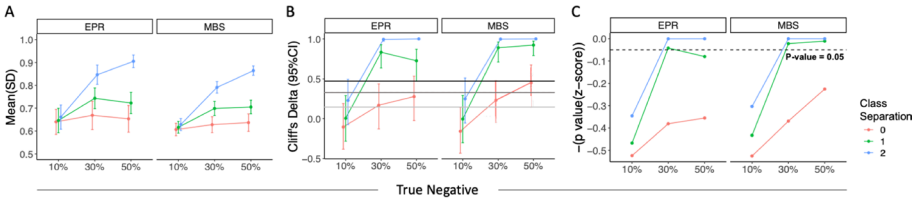

Fig. 4 Statistical methods to evaluate scores between actual and permuted label group in the PU simulations from synthetic datasets. Explicit positive recall (EPR, left) and Mean Bagging Scores (MBS, right) for varying classification difficulties. Lines are colored according to class separation distance between U and P classes and x-axis indicates the composition of the unlabeled class (% True Negatives). **A** Mean and standard deviation of the scores in actual group obtained from 30-time repetition. **B** Cliff's Delta estimate between scores from the actual P/U label and multiple sets of permuted labels. Error bars represent the 95% confidence interval of the estimate. From bottom (light gray) to top (black), lines represent the boundary of *negligible/small/medium/large* differences between groups defined by Cliff' Delta statistics. **C** Statistical significance as defined by one-tailed z-score. Dashed line indicates p value = 0.05. Lines are colored according to the underlying ground truth class separation

actual positives were not significantly (z-score p value, Fig. 4C) or substantially (Cliff's Delta, Fig. 4B) different than those observed when labels were permuted.

Evaluating confidence in class label inferences based on real-world data

Given results with synthetic data sets, we next applied this approach to actual biomedical data sets. Using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset (Fig. 3A, bottom), we observed confident differences between actual known positive and permuted positive samples for all proportions of true negatives (Fig. 3B, bottom). In this dataset, even when EPR and MBS scores approached 0.5, the U-AUC values consistently surpassed 0.9. Additionally, statistical significance was observed across all analyzed metrics when comparing scores generated under actual and permuted P/U labels.

We further broadened our analysis to encompass additional real-world datasets (Fig. 5). In this extension, we also varied the percentage of randomly selected known positive samples ranging from as few as 10% to as many as 40% (Fig. 5A). This variation allowed us to examine the impact of a low proportion of known positives among all underlying positive samples and to explore the potential bias resulting from the aggregation of the bootstrapped classifier with a reduced number of training samples. To simulate the P/U learning scenario, we randomly selected a varying number of known positives (N_{KP}) from the underlying true positives in multiple real-world datasets, including the WDBC, TCGA, and Lakhashe et al. study datasets.

In each of the datasets tested, high U-AUC values were observed (Fig. 5A, B), and actual known positive subject EPR and MBS values were statistically significantly different than observed for permuted positive subjects, regardless of the number of known positives (Fig. 5A). For WDBC and BRAC/LUAC, the largest data sets, increasing N_{KP} had a limited effect on permuted positive class scores, though it led to increasing values for the actual positive class label scores. For Lakhashe et al., the smallest dataset evaluated, the distributions remained quite distinct (Fig. 5A, C) and a large effect size was observed (Fig. 5D), though increasing N_{KP} led to somewhat greater gains for permuted than actual data (Fig. 5A). In sum, EPR and MBS values obtained from the actual positive class labels, even in the context of low numbers of known positive samples, were clearly distinct from results when positive labels were permuted. Concordantly, high

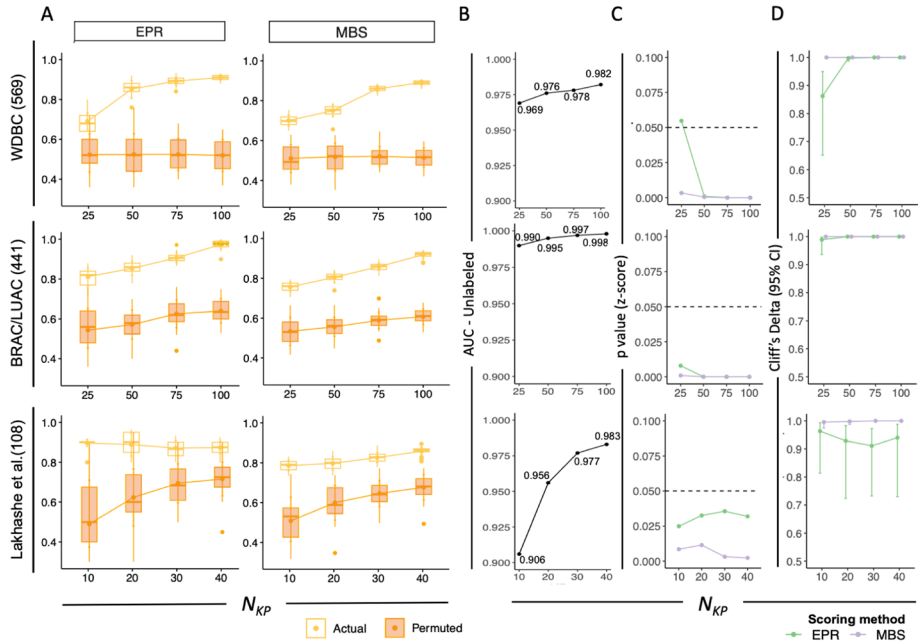

Fig. 5 Statistical methods to evaluate scores between actual and permuted label group in real world biological datasets. **A** Scores obtained from actual (yellow) and permuted (orange) P/U labels with two different scoring methods (EPR, left and MBS, right) under varied numbers of KP (N_{KP}) samples for WDBC (top) BRAC/LUAC (middle) and Lakhashe et al. (bottom) study datasets. Number of samples is indicated in parentheses. Boxplots depict median (bar), mean (point), interquartile range (IQR) and error bars depict mean and standard deviation (SD). **B** AUC of U set samples calculated between class 1 probability using PU bagging SVM compared and ground truth label. **C** Statistical significance from z-score between the mean score in actual label group and the distribution for scores in permuted label group. Dashed line: p value = 0.05. **D** Effect size (Cliff's Delta) estimate between score distributions in actual and permuted group labels

U-AUC values, significant z-score p values, and large effect sizes (Cliff's Delta) provided further evidence that comparisons between actual and permuted positive class samples can provide confidence in model results in the absence of underlying ground truth label for validation (Supplemental Table S1).

Permutation repetition to characterize the robustness of comparison metrics

Computational expense is a disadvantage of permutation testing in real-world applications, as it increases computational cost by a factor equal to the number of permuted label sets analyzed to confidently defined the null result distribution. To learn about the potential of wider use of the permutation test under the proposed methodology to evaluate the P set, we conducted a further investigation on the impact of the estimate from group comparison methods under different arbitrary choices of permutation repetition. Here, we specifically employed the Lakhashe et al. study dataset, a real-world humoral immune response dataset from a vaccine trial with a limited sample size and wide feature space (p > n). We evaluated three different levels of permutation repetitions (30, 100, 500) and four different numbers of known positive examples to evaluate the changes in statistical significance between scores from actual labels and permuted labels (Fig. 6A). For all conditions, both EPR and MBS metrics indicated a large effect size (Cliff's Delta) and high confidence (z-score p values). Compared to scores using the MBS value, a

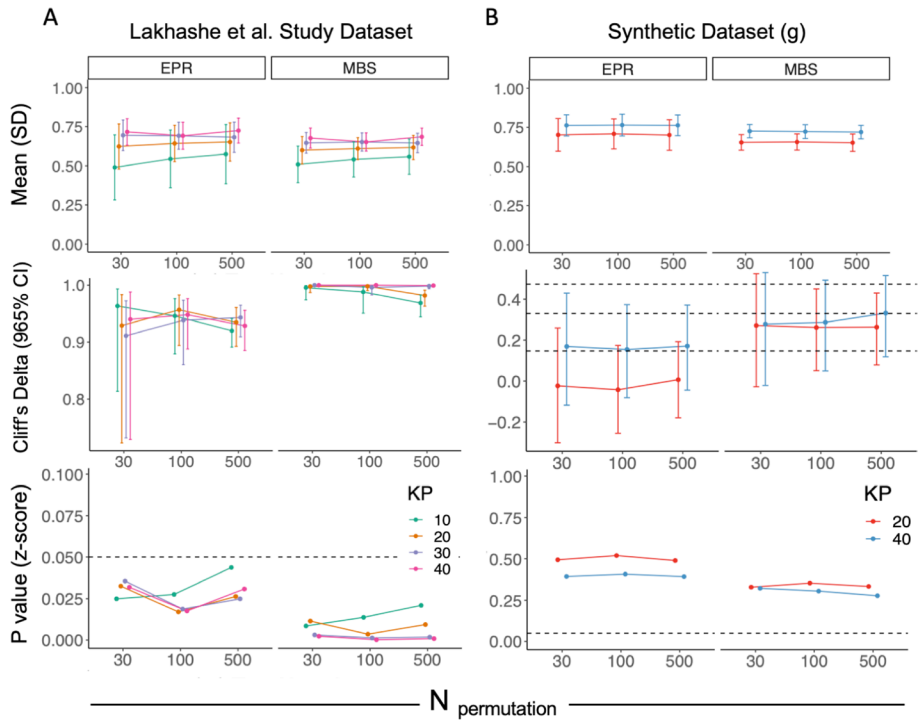

Fig. 6 *P* value from the z-score remains stable regardless of the number of permutations. Statistical comparisons between scores from actual and permuted labels in Lakhashe et al. dataset (**A**) and a dimension-matched synthetic dataset (**B**) with varied numbers of permutation repetitions. First row: sample mean and standard deviation (error bar) of the distribution of scores under the permuted labels. Second row: Cliff's Delta estimates with a 95% confidence interval. Third row: p-value from one-sample z-score. Dashed line: *p* value = 0.05

larger 95% confidence interval of Cliff's Delta was observed under a lower number of permutations; however, the group difference level remained stable over the boundary of "large" effect sizes (Cliff's Delta estimate > 0.474), and a substantial change in *p* value was not identified under different numbers of permutations in the *p* value from the z-score, which was used to compare the mean score from the actual label and the distribution of scores from multiple permutations.

As a complement to this case, in which a significant difference was obtained between scores under the actual label and permuted labels with all methods, we also investigated estimates from statistical comparison methods in a dataset where only moderate to low statistical significance was observed between scores from the actual and permuted labels. For this analysis, we generated a high dimensional synthetic dataset with a similar number of instances and attributes, but only a low level of hypercube distance (n:100, p:200, (%) TN:30) and poor to acceptable discrimination achieved in AUC of the unlabeled set based on the randomly selected KP from true positive class (TP) (KP = 20: U-AUC = 0.696; KP = 40: U-AUC = 0.733). With a 95% confidence interval of falling into the range from "negligible" to "medium" Cliff's Delta, varying the number of permutations from low to high did not positively or negatively impact the estimate from this non-parametric test (Fig. 6B). Additionally, z-score p-values also demonstrated stability

under both larger and smaller group differences regardless of the number of permutations (Fig. 6B).

Discussion

One of the primary challenges in PU learning revolves around validating binary predictions. A key point of contention is the usage of "positive" and "unlabeled" labels as substitutes for feature selection and hyperparameter tuning, given the absence of negative examples for validation. In transductive PU learning, the goal is to estimate a score function to rank the unlabeled data according to their binary class probability. One of the major limitations and concerns for both PU learning and other methods to classify unlabeled data alike is the lack of underlying ground truth labels in real-world settings to understand whether the proposed classifier is on the right track. Without setting a negative control method, it is hard to have a sense of how concordant this estimate is to underlying ground truth labels. In another word, the unlabeled samples get classified anyway. To address this dilemma, our study introduces a novel methodology that leverages and enhances traditional permutation tests within the context of PU learning. This approach provides a statistical interpretation of model confidence by comparing scores obtained from the actual P/U label with the score distribution generated from the same pipeline but with permuted P/U labels.

As we show in the synthetic data setting, how well this estimator performs depends strongly on how well separated the two classes are from each other and how representative the positive examples are compared to underlying true negative class, both in terms of quality and quantity. When this separation is close to 0, a lack of statistically significant differences can be seen between using true P/U label or permuted P/U label. These results therefore help us establish permutation testing as a 'gatekeeper' and negative control that can be used to exclude the application of PU learning from estimating unlabeled samples in datasets with small or no inherent separation—serving as a means to flag models that likely lack the ability to meaningfully classify U set samples.

To thoroughly understand the empirical behavior of the proposed methodology, we generated a diversity of scenarios, including underlying ground truth class separation, TP/TN label imbalance, and low percentages and numbers of KP samples. Here, the results from both synthetic datasets and benchmark and real-world exemplary use case datasets demonstrated that the methodology can identify good and poor prediction performance in the U set based on statistical significance and effect size obtained between scores from actual and permuted labels. To support its application under high dimensionality, which is a universal challenge in biological research datasets, we further investigated the impact of the number of permutations on statistical significance. Additionally, this study also compared two positive set scoring methods derived from the class 1 probability of the positive samples as a "spy fold". Compared to EPR, MBS generally showed a lower standard deviation in scores from both the actual label and the permuted label, leading to improved ability to capture differences. Overall, this work establishes a means to evaluate confidence in PU learning inferences in the absence of known true negatives for model validation.

Nonetheless, limitations persist. Low to moderate statistical significance between the mean score from the actual PU label and the distribution of the scores from permutation repetition when real differences existed between P and N classes (false negative results) could be generally attributed to two potential underlying issues: small separation between ground truth classes or a low proportion of true negative samples despite large separation. In the specific cases in which only 10% of true negatives existed among all the samples, we identified a rapid drop in statistical significance between positive set scores from actual P/U label and permutation repetitions, even though acceptable to excellent discrimination was suggested by the U-AUC calculated between PU bagging score of unlabeled set and ground truth. However, without a ground truth label to validate the prediction or prior knowledge of the underlying class proportion, the methodology is not yet capable of pinpointing the exact difficulty that might inflate the risk of misclassification. As a concrete example, use of this approach to predicting some classes of protein–protein interactions (PPI) may overlook genuine interactions based the scarcity of true positive examples. Existing examples (true positives) may also be biased, such as being more easily crystalized, in this example, which may also increase the likelihood of false negatives.

While we did not observe false positive results in this study, we note that use of class inferences in the setting of vaccine efficacy trials did result in the identification of false positive correlates of protection [17]. We would likewise expect in the prior PPI example, that while the class of identified as interacting proteins would be enriched in true PPI, there would almost certainly be individual members of that class that were not predicted correctly.

One of the potential explanations for the poor statistical significance despite separable ground truth classes and high expected prediction performance could be a joint effect from difficulty in classifying unlabeled samples with both few known positives and few true negatives, and limited appropriateness of permutation tests in settings with extreme imbalance. Not only was a rapid decrease in mean score from the actual P/U label associated with the former, but also the distribution of the “actual” known positive and permuted known positive tended to approximate each other in the absence of very large underlying ground truth class separation. Consequently, the current methodology still possesses the risk of yielding "false negatives" in accurately predicting the remaining unlabeled samples using the selected PU learning methods. To address this challenge, future studies could focus on improving existing PU learning algorithms to better handle extreme underlying ground truth label imbalances. Additionally, developing enhanced scoring methods for the KP set that can effectively quantify the probability distribution may prove beneficial.

Furthermore, the methodology employed in this study was evaluated using a K-fold split approach for the positive set, with K set to 5, serving as a proof-of-concept. It is worth noting that the choice of the number of fold splits can vary depending on the specific cases. Here, we hypothesized that employing a larger “K” may lead to higher computational requirements but could result in a less biased estimation of the "score" for the positive set in transductive PU learning, considering that the number of known positives (KP) generally influences the classification performance among unlabeled samples. However, we did not investigate the potential impact of “K” on the robustness reported

by permutation test in the present study, leaving room for future research to supplement this aspect. Beyond this limitation, we also only investigated a single PU learning approach. While PU bagging is frequently applied in biomedical and bioinformatic tasks, and presents advantages and limitations typical of most PU bagging methods, all methods require modelers to make some choices, and are likely to exhibit some differences in sensitivity to the factors explored here, which could impact the application and interpretation of permutation testing.

In conclusion, the effectiveness of classifying unlabeled samples using PU learning methods relies heavily on the input features provided to the PU-based classifier and the knowledge of positive examples. In this study, we introduce a permutation testing-based methodology that serves as a "gatekeeper" to assess whether the known positive samples can achieve high prediction performance in the U set. By employing a carefully selected PU-based classifier, this approach can serve as an initial step to evaluate the potential classification performance for real-world scenarios in which only positive examples are available. To the best of our knowledge, this work is the first to utilize permutation tests within the PU learning framework to establish a baseline score distribution for comparison with scores obtained from the actual labels. We emphasize the versatility of the proposed method, which demonstrates its applicability across various binary class label ratios and levels of class separation. These findings provide valuable insights for the future implementation of PU learning, particularly in high-dimensional PU learning tasks.

Methods and materials

Datasets for evaluation

The datasets that were used for this study included multivariate synthetic datasets, real-world benchmark datasets from biological and biomedical studies outside the domain of vaccinology, and one real-world humoral immune response profile from a vaccine efficacy trial in nonhuman primates (Table 1).

Synthetic datasets

For each synthetic dataset, profiles modeling responses for a set of 200 features for each of 200 subjects were generated to study the applicability of the proposed method in a dataset with limited sample size ($p \geq n$) with the scikit-learn library in Python [23]. To introduce covariance, features were composed of 30% "informative" features and 70% "redundant" features, which were generated as a random linear combination of informative features. Multiple synthetic datasets were generated with varied percentages of true negative (TN) samples (10%, 30%, 50%) and varied separation (small, medium, large) between hypercubes, which were labeled either TN or true positive (TP) (Table 1, rows a–c). Furthermore, a smaller dataset comprised of 100 samples with moderate hypercube separation was generated to assess the stability of statistical test results in the context of varied numbers of permutation replicates (Table 1, row g).

Table 1 Description of the datasets and positive-unlabeled settings evaluated

	Datasets	#Instance	#Attributes	#True negative (%)	#Known positive (%)
a	Synthetic (ClassSep = 2)	200	200	100 (50)	40 (20)
				60 (30)	
				20 (10)	
b	Synthetic (ClassSep = 1)	200	200	100 (50)	40 (20)
				60 (30)	
				20 (10)	
c	Synthetic (ClassSep = 0)	200	200	100 (50)	40 (20)
				60 (30)	
				20 (10)	
d	Wisconsin Diagnostic Breast Cancer (WDBC) https://doi.org/10.24432/C5DW2B	400	32	200 (50)	80 (20)
				120 (30)	
				43 (10.8%)	
		569	32	(212) 37.3	25 (4.40)
					50 (8.79)
					75 (13.2)
					100 (17.6)
e	TCGA-BRCA/LUAD dbGaP Study Accession: phs000178 https://portal.gdc.cancer.gov/projects/TCGA-LUAD	441	20,531	141 (32.0)	25 (5.67)
					50 (11.3)
					75 (17.0)
					100 (22.7)
f	Lakhashe study dataset	108	195	36 (33.3)	10 (9.30)
					20 (18.5)
					30 (27.8)
					40 (37.0)
g	Synthetic (ClassSep = 1)	100	200	30(30.0)	20 (20)
					40 (40)

Wisconsin diagnostic breast cancer (WDBC) dataset

The WDBC dataset is comprised of 30 numeric measurements from 10 different characteristics of cell nuclei resulting from digitalized images [18]. To evaluate the proposed method in the context of varied class label ratios, three individual datasets with 400 data points from WDBC were randomly sampled from a total of 569 instances that were initially categorically labeled as benign or malignant, with 50%, 30%, and 10.8% samples as TN. For simulation purposes, samples originally labeled "Malign" were relabeled as Class 0, representing TN, and samples labeled as "Benign" were assigned Class 1, representing TP (Table 1, row d).

BRAC/LUAC dataset

Datapoints labeled with tumor type *BRCA* (TP, n = 300) and *LUAD* (TN, n = 141) were selected from the PANCAN dataset, comprised of RNA Seq gene expression [19]. To speed up the training and testing time considering the high dimensionality of RNA sequencing data feature space, Principal Component Analysis (PCA) was performed to identify the top 325 principal components, which were selected to retain 95% percent of the variance in this dataset (Table 1, row e).

Lakhashe study dataset

Humoral immune profiles of responses elicited in 36 Rhesus Macaques by one of three distinct vaccine regimens across three distinct timepoints over the series of immunizations were profiled by multiplex immunoassay [20]. Among the three immunization regimens (M, K, L), samples from group L, which displayed overall vaccine efficacy, were defined as TN (n = 36) (Table 1, row f).

Positive-unlabeled scenario simulation

Without external prior knowledge of sample weights or the distribution of known positive (KP) samples, Positive and Unlabeled (P/U) labels were assigned by randomly selecting KP from the TP class and leaving the rest of the data points as unlabeled (U). NumPy random seed functions in Python were used for the purpose of reproducing the results only [24].

Analysis pipeline

K-fold spy positive

Adapted from the concept of the "spy positive technique" in PU learning, the KP set was first randomly allocated into k folds, with one of the folds reassigned to the U set each time as a "spy fold" [25]. The PU learning method described below was then employed with the remainder of the KP samples and the updated U set (U + spy fold) to predict the Class 1 probability of all samples in the updated U set. The Bagging Scores of all "spy positive" samples were pooled. For each set of actual KP, 30 different k-fold splits were performed to define variance (Fig. 1A).

PU learning

Transductive positive-unlabeled bootstrapped aggregation (PU bagging) described by Mordelet and Vert [5] was employed as the PU method in this study. According to the number of KP in each dataset, a matched number of unlabeled subjects were randomly sampled with replacement (bootstrapped) from the U set and temporarily labeled as "negative". A classifier was built with all KPs and bootstrapped unlabeled samples to predict the remaining out-of-bag (OOB) samples in the U set as negative or positive (Class 0 or 1). To improve numerical stability, prior to transformation, initial parameters for centering and scaling features were calculated from the "Bagged" samples. The PU Bagging score for each U sample was defined by aggregation from 100-time repeated bootstrapping as follows:

$$PU\ Bagging\ Score(U\ set) = \frac{The\ sum\ of\ prediction\ while\ "OOB"}{Total\ times\ as\ "OOB"}$$

Support vector machine with the radial basis function kernel (SVM-RBF) was used as a classifier in the PU bagging approach; it combines multiple polynomial kernels to project the non-linearly separable data into higher dimension spaces to separate the targeted classes with a hyperplane created by a linear SVM. In our previous work, we showed that PU bagging with SVM-RBF classifier using the default values from Scikit-learn

SVM package in python empirically achieved higher performance than linear SVM, as well as computational efficiency in high dimension synthetic and real-world biological data, without the benefit of ground truth labels in hyperparameter tuning. To indicate the expected prediction performance of the PU learning method to classify unlabeled samples, Area under the Received Operator Characteristic Curve (ROC-AUC) for the unlabeled samples was calculated between their PU Bagging scores and ground truth labels of U set samples.

Scoring methods

Two different scoring methods were employed to evaluate the quality of model predictions. Defined by Cheng et al. [7], explicit positive recall (EPR) is defined as the proportion of known positive examples that were predicted as "positive" among all known positive samples in the validation set. EPR was and calculated by labeling positive samples as "predicted positive" when the Bagging Score exceeded 0.5. We also consider an alternative method that lacks the requirement to set an explicit decision boundary, termed mean bagging score (MBS), which is calculated as the average Class 1 probability obtained from the PU bagging classifier for positive samples.

Label permutation

Permuted labels were generated by randomly shuffling the label of KP/U multiple times (repetition = 30, repetition = 100, repetition = 500). For each replicate, the "permuted" positive set was assessed and scored with the procedure described above for "actual" KP samples to define the baseline performance of the proposed method for the datasets analyzed.

Statistics

The scores obtained from the above steps were compared between "actual" and "permuted". The distribution of scores obtained in repetition was summarized with mean and standard deviation (SD). Furthermore, observed scores under the actual label were compared to the distribution of scores under the permuted labels, using Cliff's Delta with a 95% confidence interval, and p value from z-score. In this study, the z-score was calculated as:

$$z = \frac{\mu - \mu 0}{\sigma}$$

where μ represents the mean score of positive sets from the 30-time repeated cross-validation process; $\mu 0$ represents the mean score from repeatedly permuted labels; σ is the standard deviation of scores distribution in permutation group using $n-1$ degree of freedom. An upper-tailed test was performed under the hypothesis that $\mu > \mu 0$.

Visualization

PCA plots were developed and generated with "plotly" package in Python; T test annotation and plots were created with "ggpubr" package in R (version 4.2.2) [26, 27].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05834-2.

Additional file 1.

Acknowledgements
The authors thank Natasha Kelkar for code review and annotation.

Author contributions
Investigation, coding, data visualization, writing—original draft: SX; Writing—reviewing and editing: SX and MEA; Supervision: MEA; Conceptualization and funding: MEA.

Funding
This study was supported in part by NIAID R56AI165448 and P01AI162242.

Availability of data and materials
Code developed for this analysis is available at: https://github.com/AckermanLab/Xu_et_al_PUPermutation. Data analyzed in this study is available as follows: Wisconsin Diagnostic Breast Cancer (WDBC) https://doi.org/10.24432/C5DW2B; TCGA - BRCA/LUAD dbGaP Study Accession: phs000178 (https://portal.gdc.cancer.gov/projects/TCGA-LUAD); Lakahashe et al., from the corresponding author [20].

Declarations

Ethic approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Quantitative Biomedical Sciences Program, Dartmouth College, Hanover, NH, USA. [2]Department of Microbiology and Immunology, Geisel School of Medicine at Dartmouth, Dartmouth College, Hanover, NH, USA. [3]Thayer School of Engineering, Dartmouth College, 14 Engineering Dr., Hanover, NH 03755, USA.

Received: 6 July 2023 Accepted: 10 June 2024
Published online: 19 June 2024

References

1. Köppen M. The curse of dimensionality. In: 5th online world conference on soft computing in industrial applications (WSC5). 2000.
2. Good P. Permutation tests: a practical guide to resampling methods for testing hypotheses. Berlin: Springer; 2013.
3. Ojala M, Garriga GC. Permutation tests for studying classifier performance. J Mach Learn Res. 2010;11(6):1833–63.
4. Li F, et al. Positive-unlabeled learning in bioinformatics and computational biology: a brief review. Brief Bioinform. 2022;23(1):bbab461.
5. Mordelet F, Vert JP. A bagging SVM to learn from positive and unlabeled examples. Pattern Recogn Lett. 2014;37:201–9.
6. Bhardwaj N, Gerstein M, Lu H. Genome-wide sequence-based prediction of peripheral proteins using a novel semi-supervised learning technique. BMC Bioinform. 2010;11(1):1–8.
7. Cheng Z, Zhou S, Guan J. Computationally predicting protein-RNA interactions using only positive and unlabeled examples. J Bioinform Comput Biol. 2015;13(03):1541005.
8. Liu B, et al. Partially supervised classification of text documents. In ICML. Sydney; 2002.
9. Bekker J, Davis J. Learning from positive and unlabeled data: a survey. Mach Learn. 2020;109(4):719–60.
10. Ren Y, Ji D, Zhang H. Positive unlabeled learning for deceptive reviews detection. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.
11. Li H, et al. Spotting fake reviews via collective positive-unlabeled learning. In: 2014 IEEE international conference on data mining. 2014.
12. Yu H, Han J, Chang K-C. PEBL: Web page classification without negative examples. IEEE Trans Knowl Data Eng. 2004;16(1):70–81.
13. Cerulo L, Elkan C, Ceccarelli M. Learning gene regulatory networks from only positive and unlabeled data. BMC Bioinform. 2010;11(1):228.
14. Ward G, et al. Presence-only data and the EM algorithm. Biometrics. 2009;65(2):554–63.
15. Nagaya M, Ukita N. Embryo grading with unreliable labels due to chromosome abnormalities by regularized PU learning with ranking. IEEE Trans Med Imaging. 2022;41(2):320–31.
16. Kılıç C, Tan M. Positive unlabeled learning for deriving protein interaction networks. Netw Model Anal Health Inform Bioinform. 2012;1(3):87–102.

17. Xu S, Kelkar NS, Ackerman ME. Positive-unlabeled learning to infer protection status and identify correlates in vaccine efficacy field trials. iScience. 2024;27(3):109086.
18. Street WN, Wolberg WH, Mangasarian OL. Nuclear feature extraction for breast tumor diagnosis. In: Biomedical image processing and biomedical visualization. SPIE. 1993.
19. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The cancer genome atlas pan-cancer analysis project. Nat Genet. 2013;45(10):1113–20.
20. Lakhashe SK, et al. Cooperation between systemic and mucosal antibodies induced by virosomal vaccines targeting HIV-1 Env: protection of Indian rhesus macaques against low-dose intravaginal SHIV challenges. Front Immunol. 2022;13:788619.
21. Li X, Liu B. Learning to classify texts using positive and unlabeled data. In: IJCAI. Citeseer. 2003.
22. Hosmer DW Jr, Lemeshow S, Sturdivant RX. Applied logistic regression, vol. 398. New York: Wiley; 2013.
23. Pedregosa F, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.
24. Harris CR, et al. Array programming with NumPy. Nature. 2020;585(7825):357–62.
25. Liu B, et al. Building text classifiers using positive and unlabeled examples. In: Third IEEE international conference on data mining. IEEE. 2003.
26. Kassambara A, Kassambara MA. Package 'ggpubr'. R package version 0.1, 2020. 6(0).
27. Inc. P.T. Collaborative data science. 2015. https://plot.ly.

Publisher's Note