Learning Classifiers from Only Positive and Unlabeled Data

Charles Elkan et.al.

KDD'08

Notation

- Let x be an example
- Let $y \in \{0,1\}$ be a binary label
- Let $s = 1$ if the example x is labeled, and let $s = 0$ if x is unlabeled

$s = 1 \rightarrow y = 1$ Only positive examples are labeled.

- View x, y and s as random variables
- There is some fixed unknown overall distribution $p(x, y, s)$ over triples $\langle x, y, s \rangle$
- Training set: unlabeled examples $\langle x, s = 0 \rangle$
  labeled examples $\langle x, s = 1 \rangle$

- Goal: learn a function $f(x) = p(y = 1|x)$

Central result

- Basic assumption: the labeled positive examples are chosen completely randomly from all positive examples.

Lemma 1: Suppose the "selected completely at random" assumption holds. Then $p(y = 1|x) = p(s = 1|x)/c$ where $c = p(s = 1|y = 1)$.

$$p(+|x) * p(L|+) = p(L|x)$$

c

要求的f

样本被标记的概率，
可以观察到，记为g

Estimate c

- Labeled vs. unlabeled : $g(x) = p(s = 1|x)$
- $f * c = g$

- Let V be such a validation set that is drawn from the overall distribution $p(x, y, s)$
- Let P be the subset of examples in V that are labeled
- Estimator of c:
  $c_e = \frac{1}{n}\sum_{x \in P} g(x)$, where n is the cardinality of P.

$$p(L|x) * p(+) = p(L|+)$$

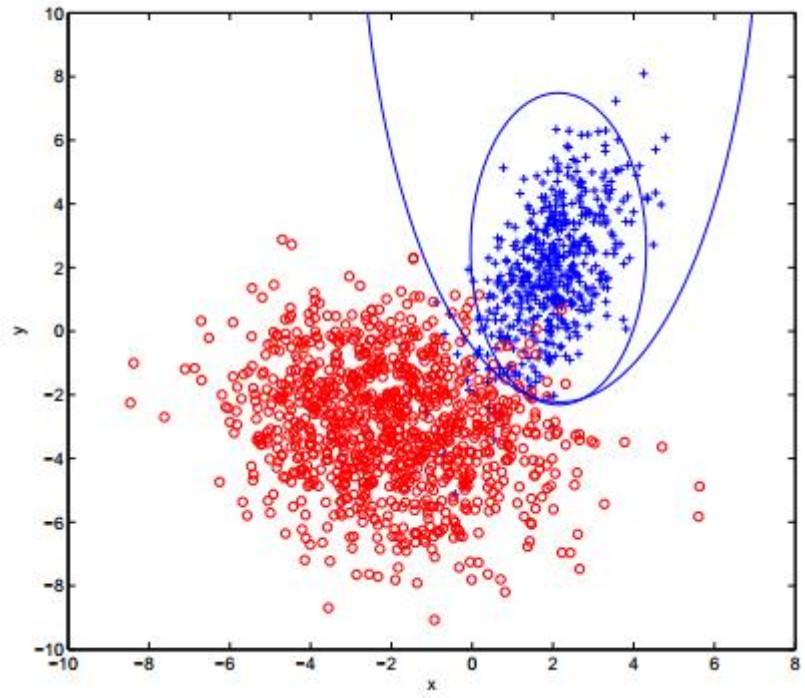
Illustration

- Generate 500 positive data points and 1000 negative data points, each from a two-dimensional Gaussian.

- Train two classifiers:
  - one using all the data
  - one using 20% of the positive data as labeled positive examples, versus all other data as negative examples.

- Based on a validation set of just 20 labeled examples, this estimated value is c_e= 0.1928, which is very close to the true value 0.2.

Illustration



- Data points lie in two dimensions.
- Blue pluses are positive examples, while red circles are negative examples .

- The large ellipse is the result of logistic regression trained on all the data.
- The small ellipse is the result of logistic regression trained on positive labeled data versus all other data, then transformed following Lemma 1.

Weighting unlabeled examples

- Let the goal be to estimate $E_{p(x,y,s)}[h(x,y)]$ for any function h, write as $E[h]$

$$
\begin{aligned}
E[h] &= \int_{x,y,s} h(x,y)p(x,y,s) \\
&= \int_x p(x) \sum_{s=0}^{1} p(s|x) \sum_{y=0}^{1} p(y|x,s)h(x,y) \\
&= \int_x p(x) \Big(p(s=1|x)h(x,1) \\
&\qquad + p(s=0|x)[p(y=1|x,s=0)h(x,1) \\
&\qquad\qquad + p(y=0|x,s=0)h(x,0)] \Big).
\end{aligned}
$$

- Less obviously , $p(y=1|x,s=0) = \frac{1-c}{c} \frac{p(s=1|x)}{1-p(s=1|x)}$

Weighting unlabeled examples

- Estimate of $E[h]$ is then the empirical average

$$\frac{1}{m}\Big(\sum_{\langle x,s=1\rangle} h(x,1) + \sum_{\langle x,s=0\rangle} w(x)h(x,1) + (1-w(x))h(x,0)\Big)$$

where

$$w(x) = p(y=1|x,s=0) = \frac{1-c}{c}\frac{p(s=1|x)}{1-p(s=1|x)} \qquad (3)$$

m is the cardinality of the training set

- Each labeled example is treated as a positive example with unit weight
- Each unlabeled example is treated as a combination of a positive example with weight $p(y=1|x,s=0)$ and a negative example with complementary weight $1-p(y=1|x,s=0)$

Application To Real-World Data

- P: 2453 records obtained from *TCDB*
- U: 4906 records selected randomly from *SwissProt* excluding its intersection with *TCDB*

- Compare four approaches(Q is set of unlabeled positive examples):

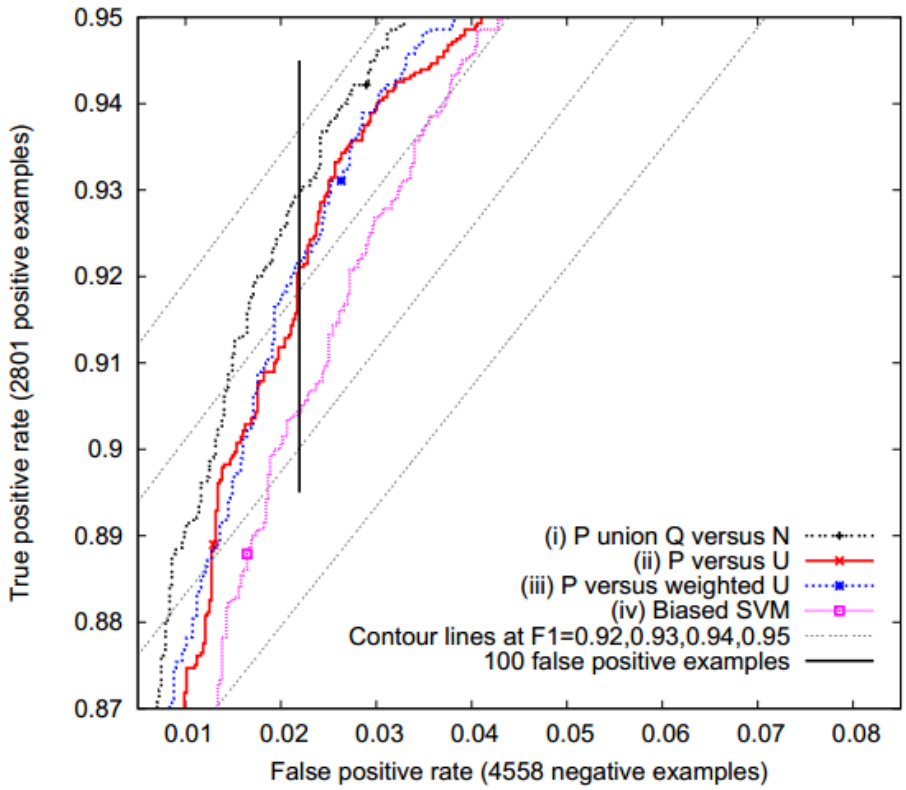
(1) standard learning from $P \cup Q$ versus N
(2) learning from P versus U with adjustment of output probabilities
(3) learning from P and U after double weighting of U
(4) the biased SVM method: make losses on P be penalized more heavily than losses on U

$$\text{minimize } \frac{1}{2}||w||^2 + C_P \sum_{i \in P} z_i + C_U \sum_{i \in U} z_i$$

$$\text{subject to } y_i(w \cdot x + b) \geq 1 - z_i \text{ and } z_i \geq 0 \text{ for all } i.$$

$$C_U = 0.01, 0.03, 0.05, \ldots, 0.61$$
$$C_P/C_U = 10, 20, 30, \ldots, 200$$

Experiment



- In order to show the differences between methods better, only the important part of the ROC space is shown.

Comparison

Table 1: Measures of performance for each of four methods.

method	accuracy	F1 score	area under ROC curve	relative time
(i) Ideal: Training on $P \cup Q$ versus N	0.9600	0.9472	0.9912	1
(ii) Training on P versus U	0.9497	0.9308	0.9895	**1**
(iii) Training on P versus weighted U	**0.9568**	**0.9422**	**0.9899**	2
(iv) Biased SVM	0.9465	0.9279	0.9895	621

Contributions

- $f = g/c$
- double weighting for U

Shortcoming :

- Strong assumption: The labeled positive examples are chosen completely randomly from all positive examples
- The estimator of c maybe not accurate

Related work

- **Cost-sensitive and Class-prior Estimation**

  Masashi Sugiyama

  - Analysis of Learning from Positive and Unlabeled Data. NIPS'14
  - Class-prior Estimation for Learning from Positive and Unlabeled Data. ACML'15

- **Scalability**

  Emanuele Sansone

  - Efficient Training for Positive Unlabeled Learning, T-PAMI 2017