

MTS Report

Oktar Ozgen

Learning classifiers without negative examples by Charles Elkan

There are approximately 1000 databases of protein examples in molecular biology. The positive examples in databases are those in which genes and proteins are included. TCBD and SwissProt are the main examples of such databases that are well studied.

The scenario is formalized in the following way. X is a vector of protein. Y is a binary value that defines the label of x and it is used to decide if x is relevant to TCBD or not. S is another binary variable that states if the label of x is known in advance or not. If $s=0$, that means we don't know if x is a positive or negative example.

There are three types of learning. First one is the one-class learning that only includes positive examples and does not include negative and unlabeled examples. Second one is semi-supervised learning. This one includes some positive and some negative and unlabeled examples. The third one is the sample solution bias. In this one, positive and negative examples are given by different distributions. Also, unlabeled examples may be available.

In term of probabilities, all data follow same fixed distributions $p(x,y,z)$ over $\langle x,y,z \rangle$. The SCAR (selected completely at random) assumption is used. This suggests that all labeled positives are chosen randomly among all positives. If $y=1$, probability of labeling is constant regardless of x . This is noted as $p(s=1|x,y=1) = p(s=1|y=1) = c$. s and x are conditionally independent given y . Non-traditional classifiers can be perceived as functions that takes x as input and returns the predicted label of x . This is noted as $p(y=1|x)$. It is possible to learn $p(s=1|x)$ from labeled and unlabeled examples.

Main lemma states that if the SCAR assumption holds, then $p(y=1|x) = p(s=1|x) / c$.

The proof of this lemma can be demonstrated as the following. $P(s=1|y=1|x) = c$.

$P(s=1|x) = p(y=1 \wedge s=1|x) = p(y=1|x).p(s=1|y=1,x) = p(y=1|x)c$.

This lemma implies that if $f=p(y=1|x)$ is used to rank examples by the chance they belong to class $y=1$, then $y=p(s=1|x)$ can be directly used instead. As a result, we can find the top 1000 proteins in SwissProt that the experts should inspect.

The constant is tried to be estimated in the following way. $c = p(s=1|y=1)$, $f = g/c \leq 1$ only if $g \leq c$. For example, $g > p(s=1|y=1)$ is impossible. There is a lemma that is used to estimate the value of the constant. This lemma is actually the main contribution of their research. So, this lemma states the following. $e = p(s=1|y=1) = c$ if $g(x) = p(s=1|x)$. (that means if the non-traditional classifier is correct) e is the estimation and $e = 1/n \sum_{x \in p} g(x)$. $n = |p|$. p is the labeled example set in V . V is the validation set from $p(x,y,s)$. Then, e is correct if $g(x)=p(s=1|x)$ precisely holds for all x . However, $g(x) \neq p(s=1|x)$ for two main reasons. First of all, g is learned from random finite training set. The other reason is that, family of models from which g is selected excludes true model.

As a result, two methods can be obtained. First method can be to learn $p(s=1|x)$, estimate c and then divide. Second method can be to train a model for $p(y=1|x)$ by estimating sufficient statistics of $p(x,y,s)$. A sufficient statistic is s mean of $E_{p(x,y,s)}[h(x,y)]$. Now, the goal becomes to estimate the expectation E for any function $h(x,y)$ when $p(x,y,s)$ is overall distributed. So, they write this as $E[h]$ and define $z = ys$, and use $p(y=1|x,z=1)$ and $p(y=1|x,z=0)$. It is shown that $p(y=1|x,z=0) = (1-c/c)(p(z=1|x)/1 - p(z=1|x))$. Then, $E[h]$ can be expressed as a 3 dimensional nested integral. Plug-in estimate of $E[h]$ is the tuning set average. At this point, intuition suggests that each labeled example is a positive example with unit weight. As a consequence, third method emerges.

The third method suggests the use of two weights for labeled examples. One of them is the positive weight that is equal to $p(y=1|x,z=0)$. The other is negative weight with $w=1-p(y=1|x,z=0)$. Estimating $p(y=1)$ is emerging as a special case. Estimating $p(y=1)$ given non-traditional training data and SCAR assumption is an open problem. All arguments apply to a classifier only if it outputs the correct probability. Some methods like logistic regression produce well calibration even with given misspecifications.

In an experiment, number of positive labeled examples (P) from TCDB is 2453 and the number of randomly selected unlabeled examples (U) taken from SwissProt is 4903. U and P are disjoint. Domain knowledge suggests 10% of U might be positive. In their research they identified the actual positives in U . 3148 members of U are positive and they call this portion Q . They define $N=U/Q$.

There are four approaches to compare. First one is the standard learning from P union Q versus N . Second approach is to learn from P vs. U and then divide by c . Third one is to learn from P and U with double weighting. The last one is the biased SVM method. They found out in their results that the third approach that they came up is actually the best of all. They are justifying this result with a graph of true positive rate against false positive rate, in which it is possible to see that their method is the best among the others.