Active Learning from Positive and Unlabeled Data

Alireza Ghasemi,

Data Mining Workshops (ICDMW), 2011 IEEE
11th International Conference on

Expected Margin Sampling

$$x_* = arg\,min_{x \in U} \left| p(+|x) - p(-|x) \right|$$

$$= arg\,min_{x \in U} \left| \frac{p(x|+)p(+)}{p(x)} - \frac{p(x|-)p(-)}{p(x)} \right|$$

$$= arg\,min_{x \in U} \left| \frac{p(x|+)p(+) - p(x|-)[1 - p(+)]}{p(x)} \right|$$

$p(x|+)$: can be computed easily from the positive (available) class samples
$p(+)$: computed from other sources of information or according to a priori knowledge about the problem

$$p(x) = p(x|+)p(+) + p(x|-)p(-)$$

$$p(x|-) = \frac{p(x) - p(x|+)p(+)}{1 - p(+)}$$

$$x_* = arg\,min_{x \in U} \left| \frac{p(x|+)p(+) - \frac{p(x) - p(x|+)p(+)}{1 - p(+)}[1 - p(+)]}{p(x)} \right|$$

$$= arg\,min_{x \in U} \left| \frac{p(x|+)p(+) - p(x) + p(x|+)p(+)}{p(x)} \right|$$

Let $a_x = \frac{p(x|+)}{p(x)}$ and $P = p(+)$ which is assumed a priori known

$$x_* = arg\,min_{x \in U} \left| 1 - 2a_x P \right|$$

To relax this assumption and count for uncertainty in *P*, we compute expected value of margin for a data sample

$$x_* = arg\,min_{x \in U}\; E_P\left\{\left|1 - 2a_x P\right|\right\}$$

$$E_P\left\{\left|1 - 2a_x P\right|\right\} = \int_0^1 \left|1 - 2a_x P\right| dP = \left(1 - a_x\right) sgn\left(\frac{1}{2} - a_x\right)$$

$$x_* = \arg\min_{x \in U} \left(1 - \frac{p(x|+)}{p(x)}\right) sgn\left(\frac{1}{2} - \frac{p(x|+)}{p(x)}\right)$$

To estimate $p(x)$ and $p(x|+)$, we can use any of the parametric or non-parametric density estimation approaches like kernel density estimation or Gaussian mixture density.

Require: Set of Positive Target Samples P, Set of Negative Outlier Samples N

Require: Set of Unlabelled Data U

repeat

  $L = P + N$

  $x_* = \arg\min_{x \in U} \left(1 - \frac{p(x|+)}{p(x)}\right) sgn\left(\frac{1}{2} - \frac{p(x|+)}{p(x)}\right)$

  Ask label of x_* from user

  if x_* is labelled as target by user **then**

    $P \leftarrow P \cup \{s\}$

  else

    $N \leftarrow N \cup \{s\}$

  end if

  Perform Learning using the new training set $L = P + N$.

until Some Stopping Condition is Met

Figure 2: Active Learning from Positive and Unlabelled Data

Entropy Based Active Learning from Positive and Unlabeled Data

$$x_* = \arg\max_{x \in U} \mathcal{H}(.|x) = -[p(+|x) \log p(+|x) + p(-|x) \log p(-|x)]$$

$$\mathcal{H} = -[a_x P \log(a_x P) + (1 - a_x P) \log(1 - a_x P)]$$

$$x_* = \arg\max_{x \in U} \mathbb{E}_P\{-[a_x P \log(a_x P) + (1 - a_x P) \log(1 - a_x P)]\}$$

$$\mathbb{E}_P\{\mathcal{H}\} = \int_0^1 -[a_x P \log(a_x P) + (1 - a_x P) \log(1 - a_x P)] dP = \frac{-a_x^2 \log(a_x) + a_x + (a_x - 1)^2 \log(1 - a_x)}{2a_x}$$

$$x_* = \arg\max_{x \in U} \frac{-a_x^2 \log(a_x) + a_x + (a_x - 1)^2 \log(1 - a_x)}{2a_x}$$

Active Learning for Multivariate Time Series Classification with Positive Unlabeled Data

Guoliang He, ICTAI 2015

For an unlabeled sample *u*, suppose its nearest positive sample is P and nearest negative sample is N. we can calculate the uncertainty of this unlabeled sample *u* by

$$\mathrm{UCT}(u) = \frac{min\{Sim(u,P), Sim(u,N)\}}{max\{Sim(u,P), Sim(u,N)\}}$$

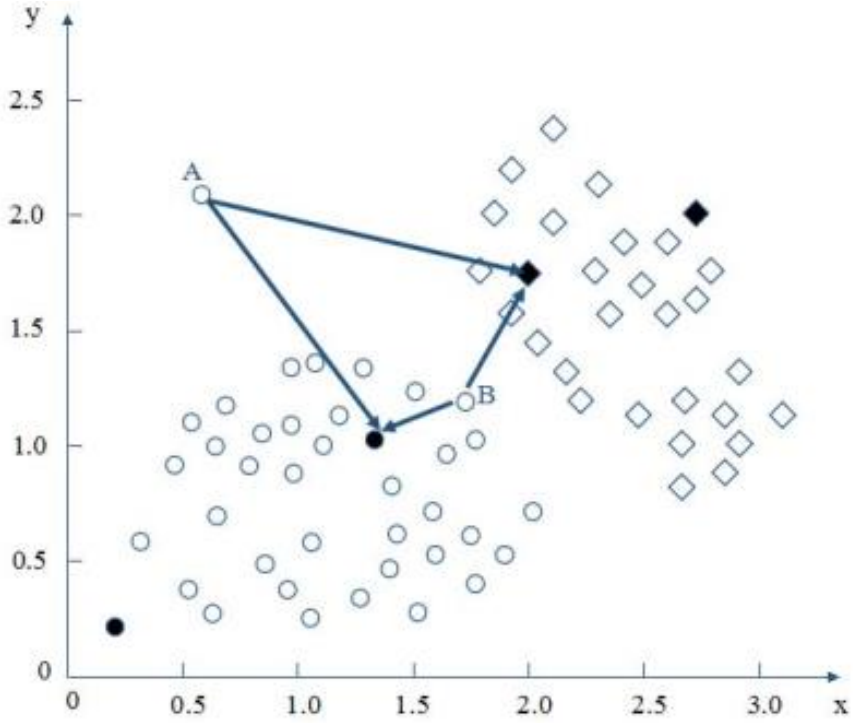

Figure 1. The utility of two unlabeled examples A and B

For an unlabeled sample A with the highest uncertainty, its utility is

$$\text{Utility}(\text{A}) = \frac{N_i}{|U| - K}$$

Where N_i is the number of neighbors of A, $|U|$is the number of the unlabeled samples and K is the number of unlabeled samples the highest uncertainty.

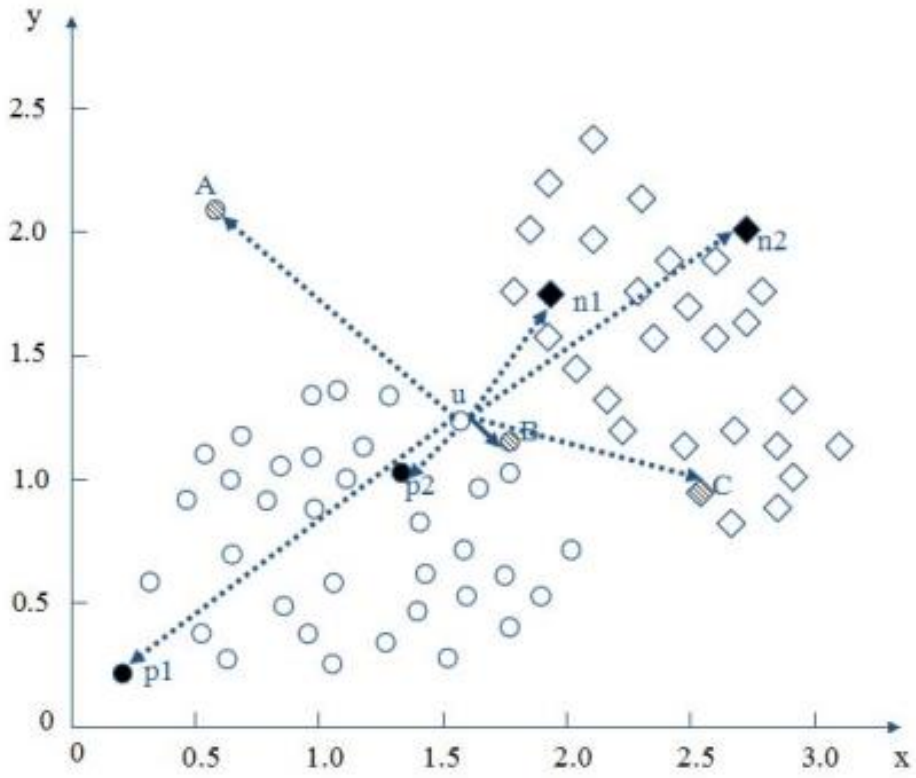

Figure 2. The profile to evaluate the utility of an unlabeled example B

Input: a training dataset with a few positive samples P and huge number unlabeled samples U
Output: the labeled dataset D
1: initialize D = P
2: labeling a confident negative sample u in unlabeled data U
4: D = P + {u}; U= U- {u}
3: Do
4: Selecting an optimal unlabeled sample x within U
5: Labeling x manually
6: D = P + {x}; U= U- {x}
7: While (stopping condition is not satisfied)
8: Return D