随着正例样本的增多，U中的正类样本会越来越少。这样一来会破坏U的原始数据分布，不满足PU Learning的基本假设:

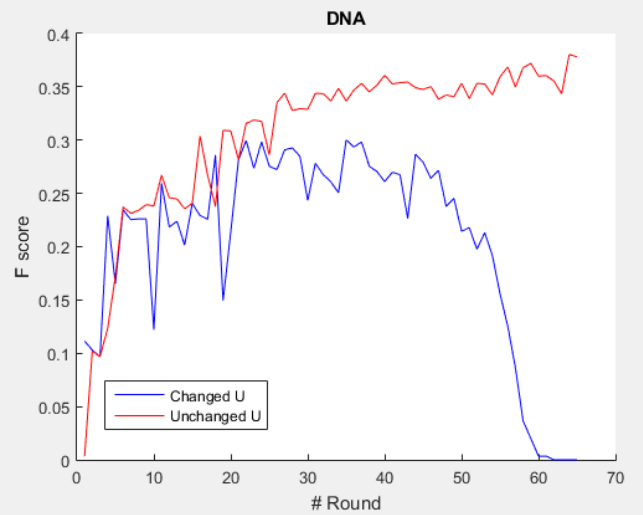
$$\mathcal{X}_{\mathrm{P}} := \{\boldsymbol{x}_i^{\mathrm{P}}\}_{i=1}^{n_{\mathrm{P}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{P}}(\boldsymbol{x}) := p(\boldsymbol{x} \mid y = +1),$$

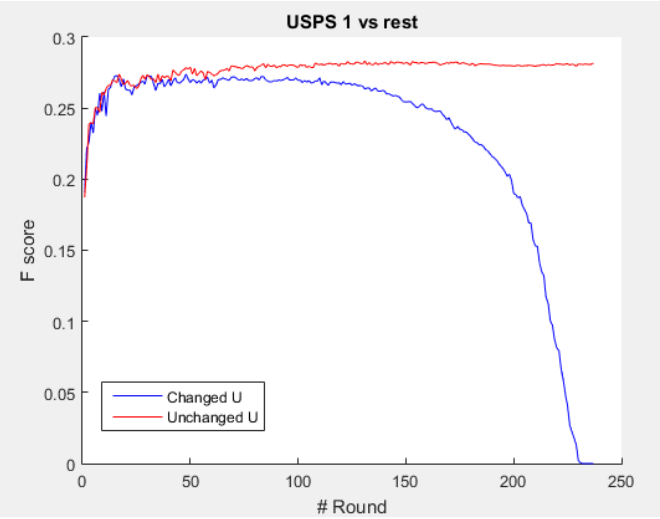
$$\mathcal{X}_{\mathrm{U}} := \{\boldsymbol{x}_i^{\mathrm{U}}\}_{i=1}^{n_{\mathrm{U}}} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}) := \theta_{\mathrm{P}} p_{\mathrm{P}}(\boldsymbol{x}) + \theta_{\mathrm{N}} p_{\mathrm{N}}(\boldsymbol{x}),$$

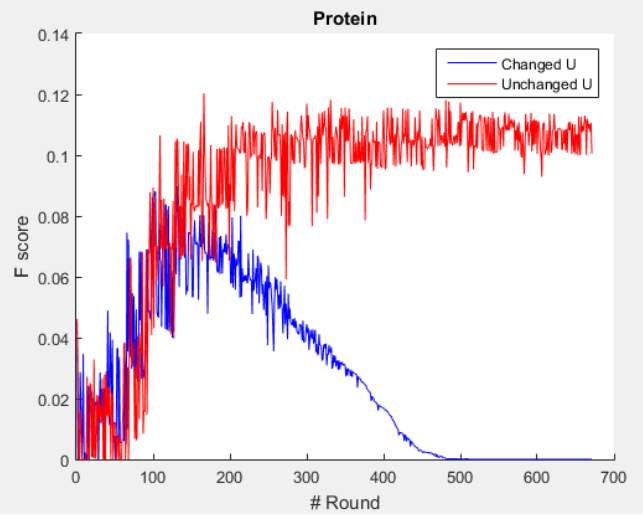
能不能只是从U中选择正类，但是保持U不改变。

Distribution of Unlabeled Data

	# labels	# instances	# positive	# features
DNA	3	2,000	331	180
USPS	10	7291	1194	256
Protein	3	17,766	3362	357







$$f(x) = w^T \phi(x)$$

$$\min_{w,Q} \widehat{R}_{P \cup Q} + \widehat{R}_U$$

$$= \min_{w,Q} - \frac{\pi}{|P \cup Q|} \sum_{x_i} f(x_i) + \frac{1}{|U|} \sum_{x_j} l(f(x_j), -1) + \frac{\lambda}{2} \|f\|_H^2$$

$$= \min_{w,Q} - \frac{\pi}{n_P + b} \sum_{x_i \in P \cup Q} w^T \phi(x_i) + \frac{1}{n_U} \sum_{x_j \in U} \frac{\left(w^T \phi(x_j) + 1\right)^2}{4} + \frac{\lambda}{2} w^T w$$

$$= \min_{q^T 1_U = b, w} - \frac{\pi}{n_P + b} \left(\sum_{x_i \in P} w^T \phi(x_i) + \sum_{x_j \in U} q_j p_j w^T \phi(x_j) \right) + \frac{1}{4 n_U} \sum_{x_j \in U} \left(w^T \phi(x_j) + 1\right)^2 + \frac{\lambda}{2} w^T w$$

- If query index q is fixed, the objective is to find the optimal classifier.
- If w is fixed, the objective is to find the max $f(x)$ in unlabeled set.

$$\min_{Q,w} -\frac{\pi}{n_P+b}\sum_{x_i\in P\cup Q} w^T\phi(x_i)+\frac{1}{n_U}\sum_{x_j\in U}\frac{\left(w^T\phi(x_j)+1\right)^2}{4}+\frac{\lambda}{2}w^Tw+MMD^2\left(P\cup Q,\hat{P}\right)$$

$$=\min_{q^T1_U=b,w} -\frac{\pi}{n_P+b}\left(\sum_{x_i\in P} w^T\phi(x_i)+\sum_{x_j\in U} q_j\left[w^T\phi(x_j)\right]^2\right)+\frac{1}{4n_U}\sum_{x_j\in U}\left(w^T\phi(x_j)+1\right)^2+\frac{\lambda}{2}w^Tw$$

$$+\left\|\frac{1}{n_P+b}\sum_{x_i\in P\cup Q}\phi(x_i)-\frac{1}{n_{\hat{P}}}\sum_{x_i\in\hat{P}}\phi(x_i)\right\|_F^2$$

MMD:

$$\left\|\frac{1}{n_p+b}\left(\sum_{x_i\in P}\phi(x_i)+\sum_{x_i\in Q}\phi(x_i)\right)-\frac{1}{n_{\hat{p}}}\sum_{x_i\in\hat{P}}\phi(x_i)\right\|_F^2$$

$$=\left\|\frac{1}{n_p+b}\left(\sum_{x_i\in P}\phi(x_i)+\sum_{x_i\in U}q_i\phi(x_i)\right)-\frac{1}{n_{\hat{p}}}\sum_{x_i\in\hat{P}}\phi(x_i)\right\|_F^2$$

$\hat{P}$、U之间有重复

$$=\frac{1}{\left(n_p+b\right)^2}\left[2K_{PU}q+q^TK_{UU}q\right]-\frac{2}{\left(n_p+b\right)n_{\hat{p}}}K_{\hat{P}U}q+const$$

examples with less similarity with labeled positive data are more likely to be selected ensuring *diversity*

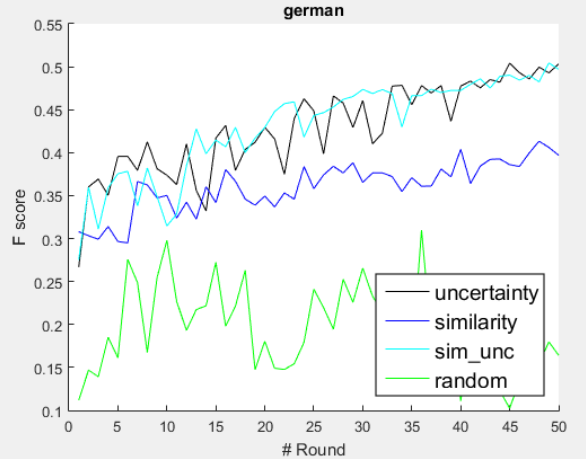
ensures that the selected query set has minimum similarity within itself, avoiding *redundancy*

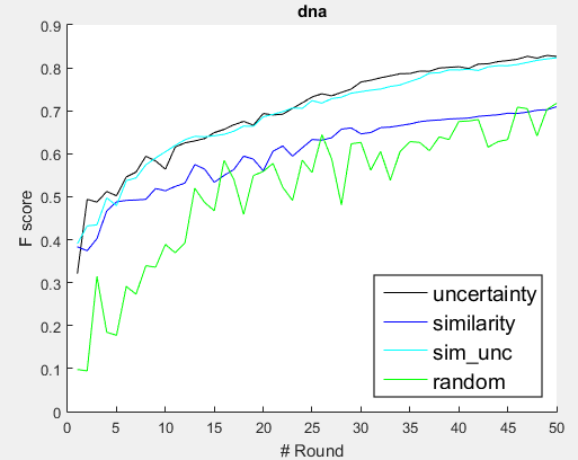
enforces the selected examples to be similar to the predicted positive samples, ensuring *positivity*

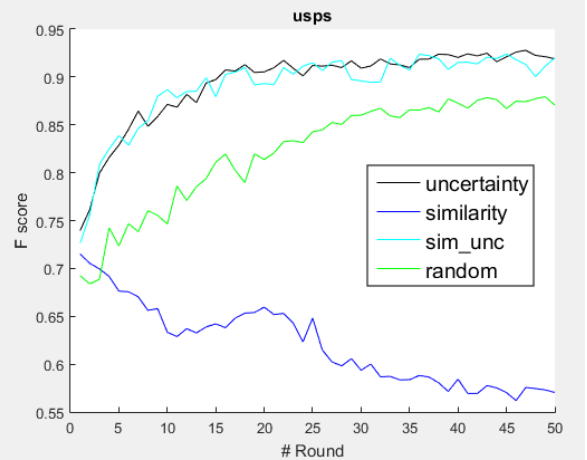
Experiment

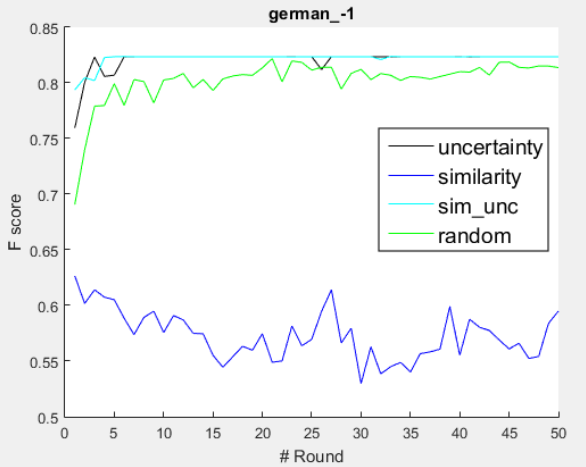
1. Random
2. Similarity: 与正类的相似程度，用$f(x)$计算，选择值最大的一批样本
3. Uncertainty: 分类器对于样本的不确定度，用$|f(x)|$计算，选择值最小的一批样本
4. Similarity – Uncertainty
5. Proposed: ERM

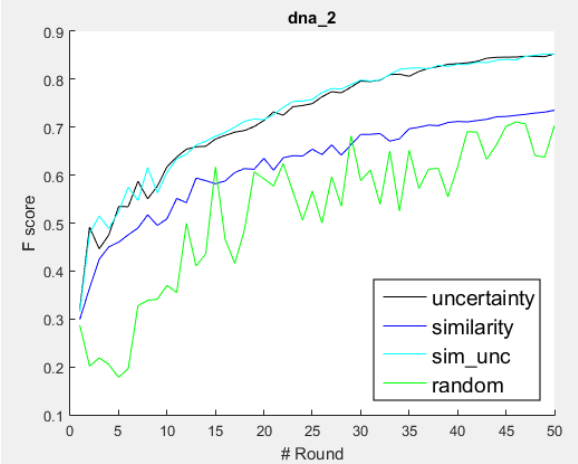
Results

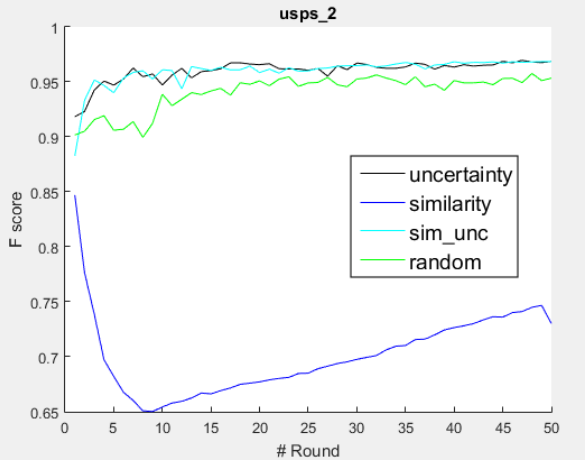

german
F score
Round
uncertainty
similarity
sim_unc
random


dna
F score
Round
uncertainty
similarity
sim_unc
random


usps
F score
Round
uncertainty
similarity
sim_unc
random


german_-1
F score
Round
uncertainty
similarity
sim_unc
random


dna_2
F score
Round
uncertainty
similarity
sim_unc
random


usps_2
F score
Round
uncertainty
similarity
sim_unc
random