数据分布的影响

D: 真实数据分布
P: 真实正类分布
Q: 正类标记数据的分布（已标记正类+Query）
U: 观测到的未标记数据分布

$$
\begin{aligned}
E_D[l(g)] &= \pi E_P[l(g,+)] + (1-\pi)E_N[l(g,-)] \\
&= \pi E_P[l(g,+)] + \{E_D[l(g,-)] - \pi E_P[l(g,-)]\} \\
&= \pi E_P[l(g,+) - l(g,-)] + E_D[l(g,-)] \\
&\le \pi \widehat{E}_P[\widetilde{l}(g)] + \widehat{E}_U[l(g,-)] + \ldots
\end{aligned}
$$

Q是P的偏分布

$$
E_D[l(g)] \le \pi\{E_P[\widetilde{l}(g)] - E_Q[\widetilde{l}(g)]\} + \pi\widehat{E}_Q[\widetilde{l}(g)] + \widehat{E}_U[l(g,-)] + \ldots
$$

Experiment

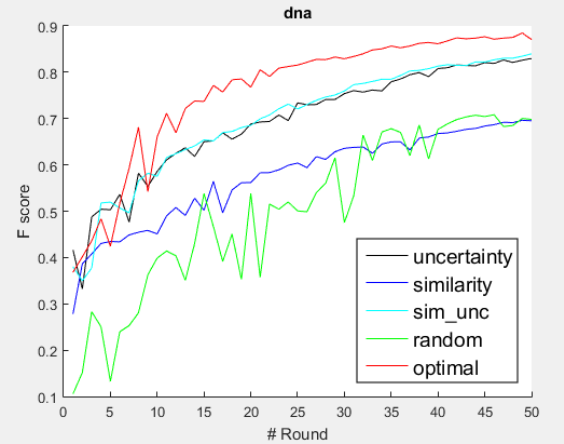
实验目的：验证已标记正类数据分布的影响

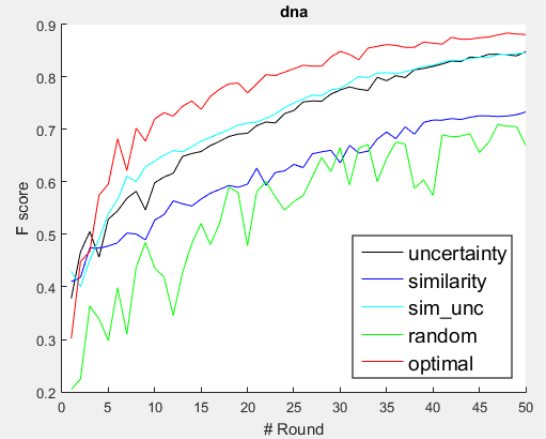
1. Random
2. Similarity: 与正类的相似程度，用$f(x)$计算，选择值最大的一批样本
3. Uncertainty: 分类器对于样本的不确定度，用$|f(x)|$计算，选择值最小的一批样本
4. Similarity – Uncertainty
5. Optimal: select positive from U randomly

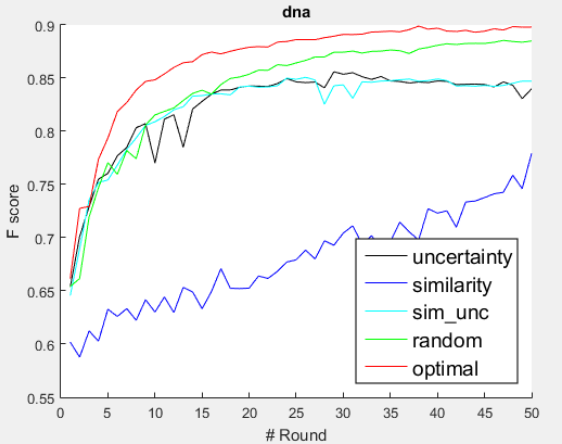
DNA

Value	Count	Percent
1	331	23.64%
2	329	23.50%
3	740	52.86%

Training: 2,000
Test: 1,186
Feature: 180



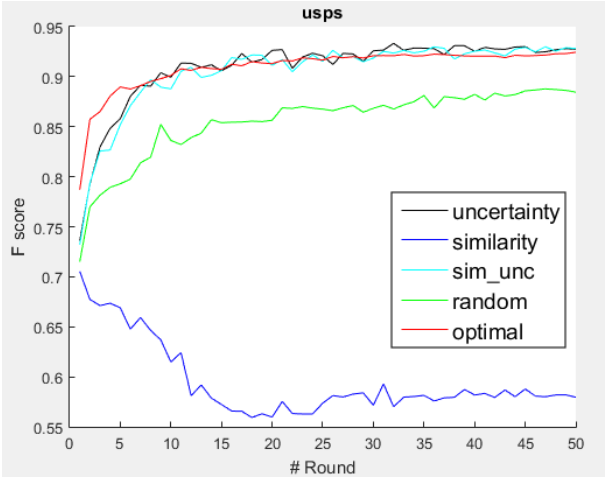


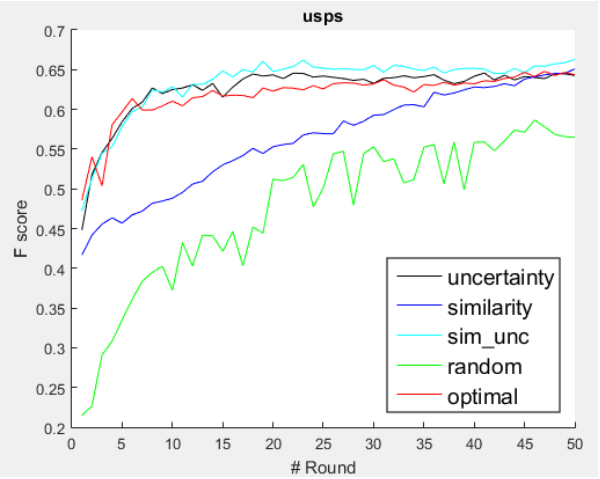


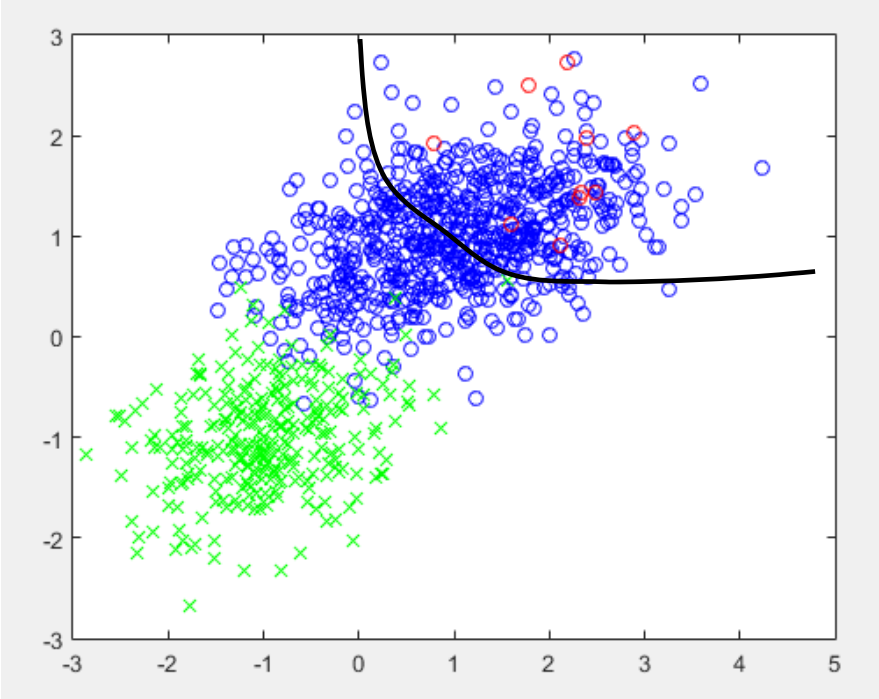
USPS

Value	Count	Percent
1	1194	16.38%
2	1005	13.78%
3	731	10.03%
4	658	9.02%
5	652	8.94%
6	556	7.63%
7	664	9.11%
8	645	8.85%
9	542	7.43%
10	644	8.83%

Training: 7,291
Test: 2,007
Feature: 256



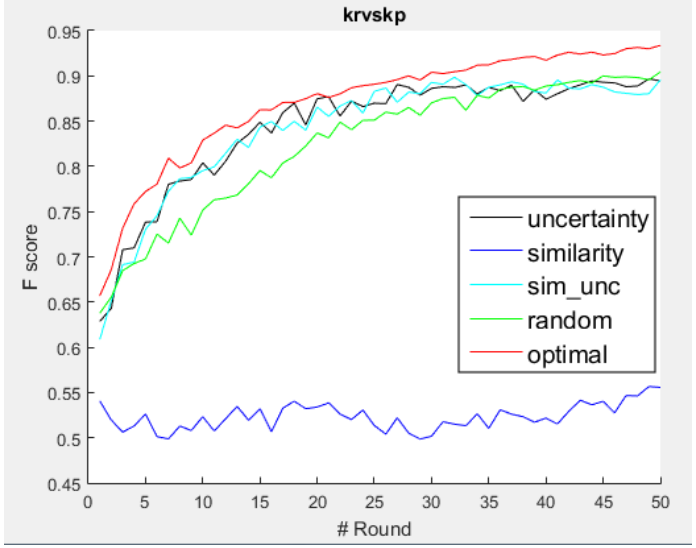


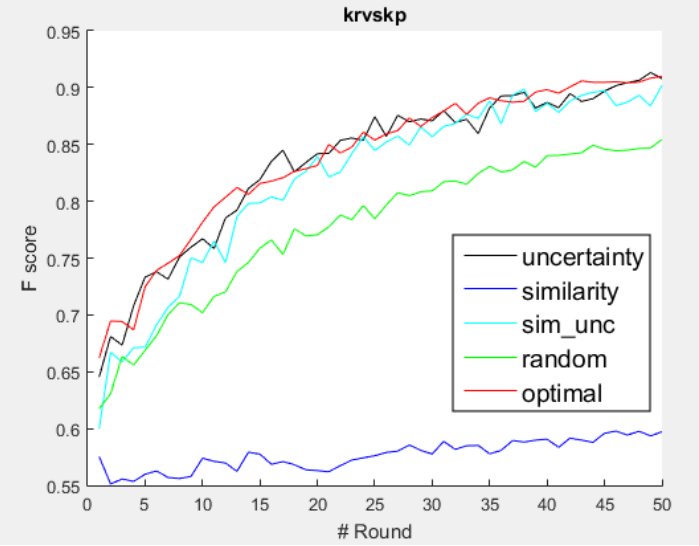

3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
4
5

Krvskp

Value	Count	Percent
-1	1168	52.21%
1	1069	47.79%

Training: 2,237
Test: 959
Feature: 36





Satimage

Value	Count	Percent
1	781	25.16%
2	384	12.37%
3	798	25.71%
4	306	9.86%
5	294	9.47%
6	541	17.43%

Training: 4,435
Test: 2,000
Feature: 36