On Discriminative Learning of Prediction Uncertainty

ICML 2019

Introduction

$$R(h) = \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \ell(y, h(x))\, dx \,,$$

where $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a *loss* penalizing the predictions

In classification with a reject option, the classifier is allowed in uncertain cases to abstain from prediction.

Cost-based model

$$R_B(h, c) = \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \left(\ell(y, h(x)) c(x) + (1 - c(x)) \varepsilon \right) dx$$

Introduction

$$h_B(x) \in \operatorname*{argmin}_{\hat{y} \in \mathcal{Y}} \sum_{y \in \mathcal{Y}} p(y \mid x)\, \ell(y, \hat{y}) \tag{1}$$

$$c_B(x) = \begin{cases} 1 & \text{if} \quad r^*(x) < \varepsilon\,, \\ \tau & \text{if} \quad r^*(x) = \varepsilon\,, \\ 0 & \text{if} \quad r^*(x) > \varepsilon\,, \end{cases} \tag{2}$$

where $r^*(x) = \min_{\hat{y} \in \mathcal{Y}} \sum_{y \in \mathcal{Y}} p(y \mid x)\, \ell(y, \hat{y})$ is the minimal class conditional risk associated to the input x, and τ is any number from the interval $[0, 1]$.

Introduction

Bounded-improvement model

$$\phi(c) = \int_{\mathcal{X}} p(x)\, c(x)\, dx$$

$$R_S(h, c) = \frac{\int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y)\, \ell(y, h(x))\, c(x)\, dx}{\phi(c)}$$

$$\max_{h,c} \phi(c) \quad \text{s.t.} \quad R_S(h, c) \leq \lambda\,, \tag{3}$$

Introduction

Theorem 1 *Let (h, c) be an optimal solution to (3). Then, (h_B, c), where h_B is the optimal Bayes classifier (1), is also optimal to (3).*

$$\max_{c \in [0,1]^{\mathcal{X}}} \int_{\mathcal{X}} p(x)c(x)dx \;\; \text{s.t.} \;\; \int_{\mathcal{X}} p(x)c(x)\overline{r}(x)dx \leq 0 \;\; \text{where } \overline{r}(x) = r(x) - \lambda \tag{4}$$

$$r(x) = \sum_{y \in \mathcal{Y}} p(y \mid x)\, \ell(y, h(x))$$

Introduction

Theorem 2 *A selection function* $c^* : \mathcal{X} \to [0,1]$ *is an optimal solution to (4) if and only if it holds*

$$\int_{\mathcal{X}_{\overline{r}(x)<b}} p(x)c^*(x)dx = \int_{\mathcal{X}_{\overline{r}(x)<b}} p(x)dx, \tag{6}$$

$$\int_{\mathcal{X}_{\overline{r}(x)=b}} p(x)c^*(x)dx = \begin{cases} -\frac{\rho(\mathcal{X}_{\overline{r}(x)<b})}{b} & \text{if} \quad b > 0\,, \\ \int_{\mathcal{X}_{\overline{r}(x)=0}} p(x)dx & \text{if} \quad b = 0\,, \end{cases} \tag{7}$$

$$\int_{\mathcal{X}_{\overline{r}(x)>b}} p(x)c^*(x)dx = 0\,, \tag{8}$$

where $\rho(\mathcal{X}') = \int_{\mathcal{X}'} p(x)\overline{r}(x)\,dx$ *is the expectation of* $\overline{r}(x)$ *restricted to inputs in* $\mathcal{X}'$*, and*

$$b = \sup\{a \mid \rho(\mathcal{X}_{\overline{r}(x)\leq a}) \leq 0\} \geq 0\,. \tag{9}$$

Introduction

Corollary 1 *Let $r : \mathcal{X} \to \mathbb{R}$ be the conditional risk (5), τ the acceptance probability given by (10) and $\gamma = b + \lambda$ the rejection threshold given by the target-risk λ and b computed by (9). Then the selection function*

$$c^*(x) = \begin{cases} 1 & \text{if} \quad r(x) < \gamma\,, \\ \tau & \text{if} \quad r(x) = \gamma\,, \\ 0 & \text{if} \quad r(x) < \gamma\,, \end{cases} \tag{11}$$

satisfies the optimality condition of Theorem 2.

$$\tau = \begin{cases} 1 & \text{if} \quad \rho(\mathcal{X}_{r(x)=\gamma}) = 0\,, \\ -\frac{\rho(\mathcal{X}_{r(x)<\gamma})}{\rho(\mathcal{X}_{r(x)=\gamma})} & \text{if} \quad \rho(\mathcal{X}_{r(x)=\gamma}) > 0\,. \end{cases}$$

$$r(x_{\pi_1}) \le r(x_{\pi_2}) \le \cdots\cdots \le r(x_{\pi_n})$$

$$s(x_{\pi_1}) \le s(x_{\pi_2}) \le \cdots\cdots \le s(x_{\pi_n})$$

Discriminative Learning of Uncertainty

Order Enforcing Loss Function

Let $\mathcal{T}_n = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} \mid i = 1, \ldots, n\}$ be a set of inputs and labels generated from n i.i.d. random variables with distribution $p(x, y)$. We define a loss function $\Delta : \mathbb{R}^n \times \mathcal{X}^n \times \mathcal{Y}^n \to \mathbb{R}_+$ as

$$\Delta(s, \mathcal{T}_n) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \ell(y_i, h(x_i)) [\![s(x_i) \leq s(x_j)]\!] ,$$

Discriminative Learning of Uncertainty

$$
\begin{aligned}
E(s) &= \int_{\mathcal{X}^n} \sum_{\boldsymbol{y}\in\mathcal{Y}^n} \prod_{i=1}^{n} p(x_i, y_i)\Delta(s, \mathcal{T}_n)d\boldsymbol{x} \\
&= \frac{1}{n^2} \int_{\mathcal{X}^n} \prod_{i=1}^{n} p(x_i) \sum_{i=1}^{n}\sum_{j=1}^{n} r(x_i)\,[\![s(x_i) \leq s(x_j)]\!]\,d\boldsymbol{x} \\
&= \frac{1}{n^2} \sum_{i=1}^{n}\sum_{j=1}^{n} \int_{\mathcal{X}} \int_{\mathcal{X}} p(x)p(z)r(x)[\![s(x) \leq s(z)]\!]dz\,dx \\
&= \int_{\mathcal{X}} p(x)\,r(x) \left(\int_{\mathcal{X}} p(z)\,[\![s(x) \leq s(z)]\!]dz \right) dx\,,
\end{aligned}
$$

Discriminative Learning of Uncertainty

Corollary 3 *Let* $s^* \in \operatorname{argmin}_{s:\mathcal{X}\to\mathbb{R}} E(s)$. *The selection function* $c^* : \mathcal{X} \to [0,1]$ *defined by*

$$c^*(x) = \begin{cases} 1 & \text{if} \quad s^*(x) < \gamma\,, \\ \tau & \text{if} \quad s^*(x) = \gamma\,, \\ 0 & \text{if} \quad s^*(x) > \gamma\,, \end{cases} \tag{17}$$

$$\text{where} \quad \gamma = \sup\left\{a \mid \rho(\mathcal{X}_{s^*(x)\le a}) \le 0\right\}, \tag{18}$$

$$\text{and} \quad \tau = \begin{cases} 1 & \text{if} \quad \rho(\mathcal{X}_{s^*(x)=\gamma}) = 0\,, \\ -\frac{\rho(\mathcal{X}_{s^*(x)<\gamma})}{\rho(\mathcal{X}_{s^*(x)=\gamma})} & \text{if} \quad \rho(\mathcal{X}_{s^*(x)=\gamma}) > 0\,, \end{cases} \tag{19}$$

fulfills conditions (6), (7) and (8) of Theorem 2, therefore it is an optimal solution to (4).

Discriminative Learning of Uncertainty

$$\hat{\Delta}(\boldsymbol{\theta}, \mathcal{T}_n) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \ell(y_i, h(x_i))\, \omega\big(s_{\boldsymbol{\theta}}(x_j) - s_{\boldsymbol{\theta}}(x_i)\big)$$

where $\omega(t) = \max\{0, 1 + t\}$

$$F(\boldsymbol{\theta}, \mathcal{T}_m) = \frac{C}{2}\|\boldsymbol{\theta}\|^2 + \frac{1}{P} \sum_{k=1}^{P} \hat{\Delta}(\boldsymbol{\theta}, \mathcal{T}^k)$$

Experiments

Baseline uncertainty measure

Logistic-regression	$1 - p_{\boldsymbol{\theta}}(h(\boldsymbol{x})\|\boldsymbol{x})$
Support Vector Machines	$\min_{y \in Y} \langle \boldsymbol{w}_y, \boldsymbol{x} \rangle + b_y$

Proposed method for uncertainty learning

$$s_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle \boldsymbol{w}_{h(\boldsymbol{x})}, \boldsymbol{\psi}(\boldsymbol{x}) \rangle + b_y$$

1. $\boldsymbol{\psi}(\boldsymbol{x}) = \boldsymbol{x}$ 2. $\boldsymbol{\psi}(\boldsymbol{x}) = (x_i \cdot x_j | (i,j) \in \{1,2,\dots,d\}^2 \wedge j \leq i)$

3.Features extracted by multi-layer perceptron trained from examples.

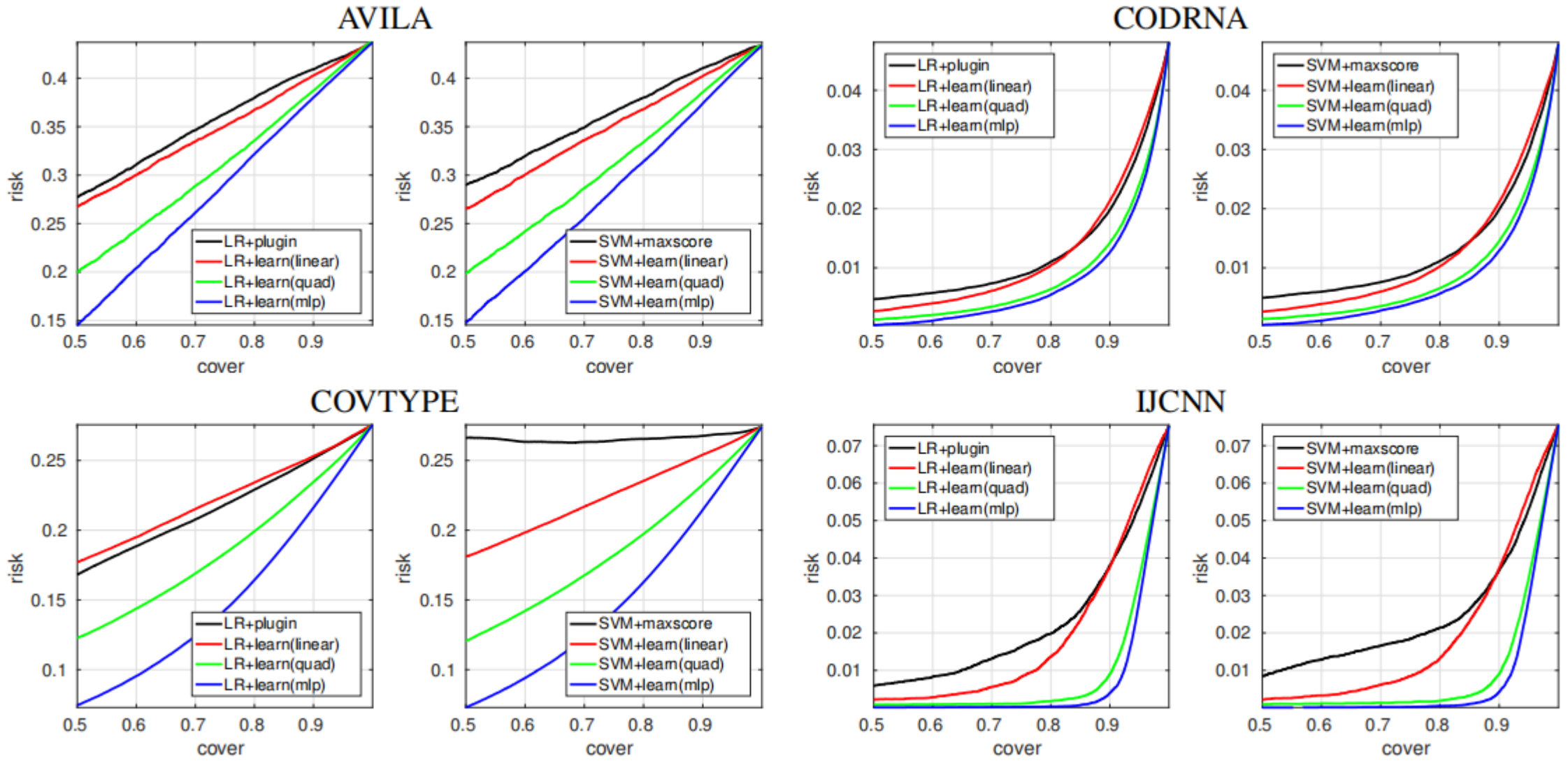
Experiments

	classifier	selection function	AUC ×100	R@90 ×100	R@100 ×100
AVILA	LR	plugin	27.2 ±0.6	40.9 ±0.5	43.7 ±0.4
	LR	learn(M)	17.3 ±0.4	38.0 ±0.4	
	SVM	maxscore	31.7 ±0.8	41.0 ±0.7	43.3 ±0.7
	SVM	learn(M)	16.9 ±0.7	37.4 ±0.8	
CODRNA	LR	plugin	0.9 ±0.0	2.0 ±0.0	4.8 ±0.1
	LR	learn(M)	0.4 ±0.0	1.3 ±0.1	
	SVM	maxscore	0.9 ±0.1	2.0 ±0.1	4.8 ±0.1
	SVM	learn(M)	0.4 ±0.0	1.3 ±0.1	
COVTYPE	LR	plugin	16.5 ±0.1	25.1 ±0.2	27.6 ±0.2
	LR	learn(M)	10.0 ±0.1	21.6 ±0.2	
	SVM	maxscore	25.7 ±0.8	26.8 ±0.1	27.4 ±0.1
	SVM	learn(M)	9.8 ±0.1	21.5 ±0.1	

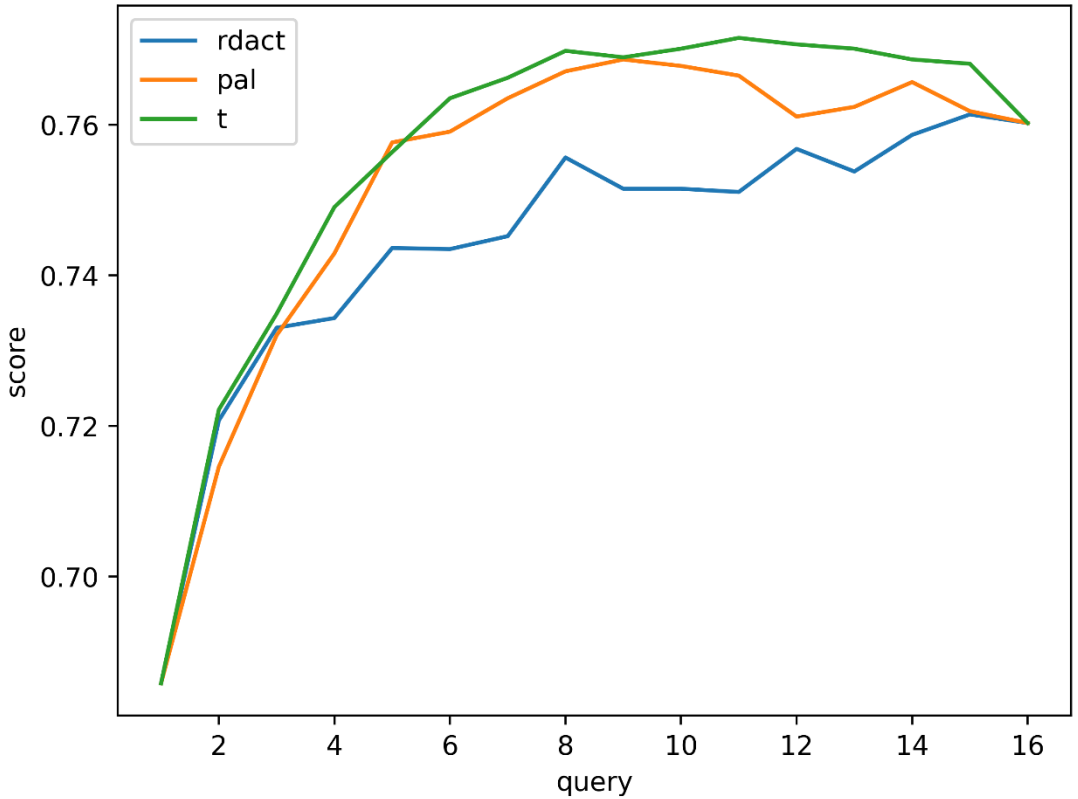
$$\hat{R}_S(i,s) = \frac{1}{i}\sum_{j=1}^{i} \ell(y_{\pi(j)}, h(x_{\pi(j)}))$$

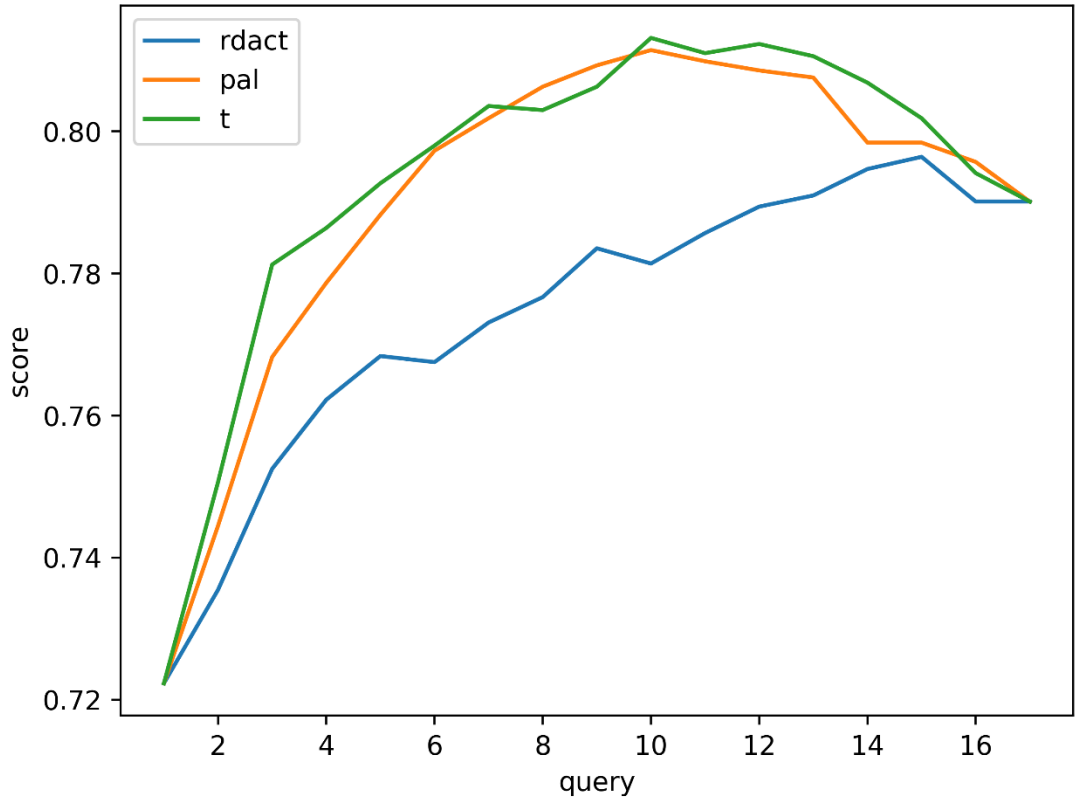
IJCNN	LR	plugin	1.3 ±0.0	3.8 ±0.1	7.5 ±0.1
	LR	learn(M)	0.3 ±0.0	0.4 ±0.1	
	SVM	maxscore	1.4 ±0.0	3.7 ±0.2	7.6 ±0.2
	SVM	learn(M)	0.4 ±0.0	0.4 ±0.1	
LETTER	LR	plugin	7.4 ±0.4	18.3 ±0.6	23.3 ±0.6
	LR	learn(Q)	4.1 ±0.1	15.4 ±0.6	
	SVM	maxscore	10.2 ±0.2	19.1 ±0.4	22.1 ±0.7
	SVM	learn(Q)	3.9 ±0.3	14.0 ±0.8	
PENDIGI	LR	plugin	0.7 ±0.0	1.9 ±0.1	5.3 ±0.4
	LR	learn(Q)	0.7 ±0.1	0.8 ±0.2	
	SVM	maxscore	2.8 ±0.4	3.9 ±0.4	4.9 ±0.6
	SVM	learn(Q)	0.8 ±0.1	0.7 ±0.2	

Experiments

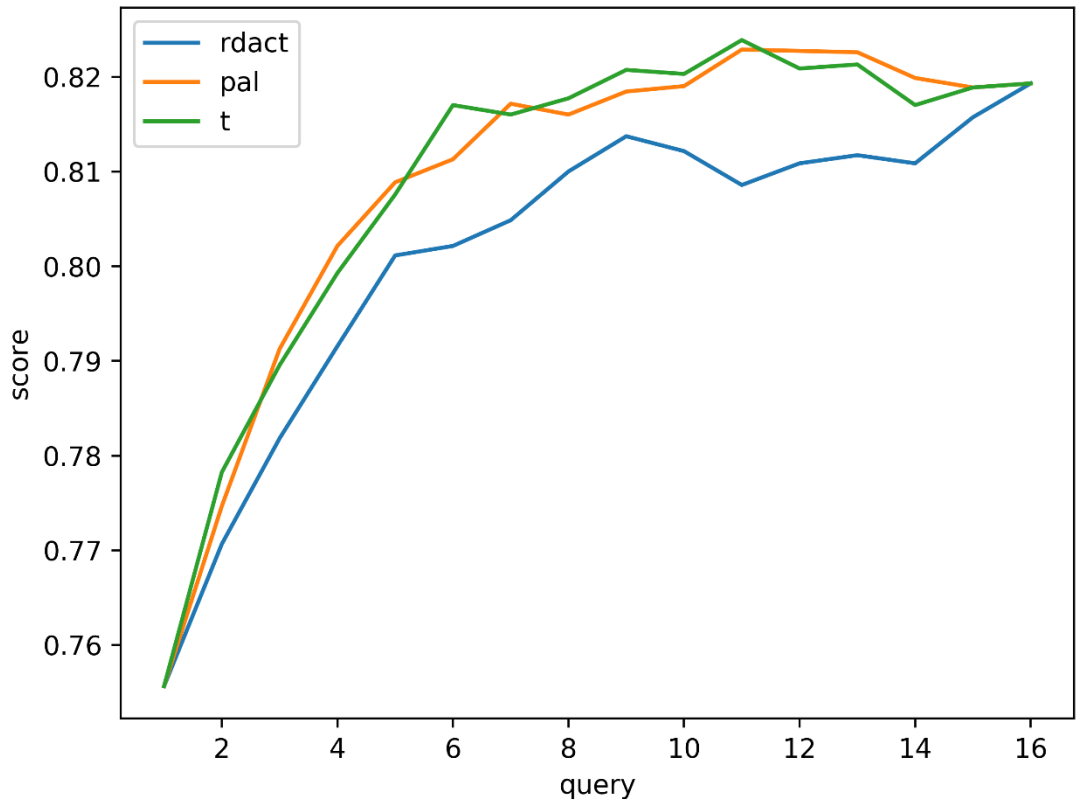


后验

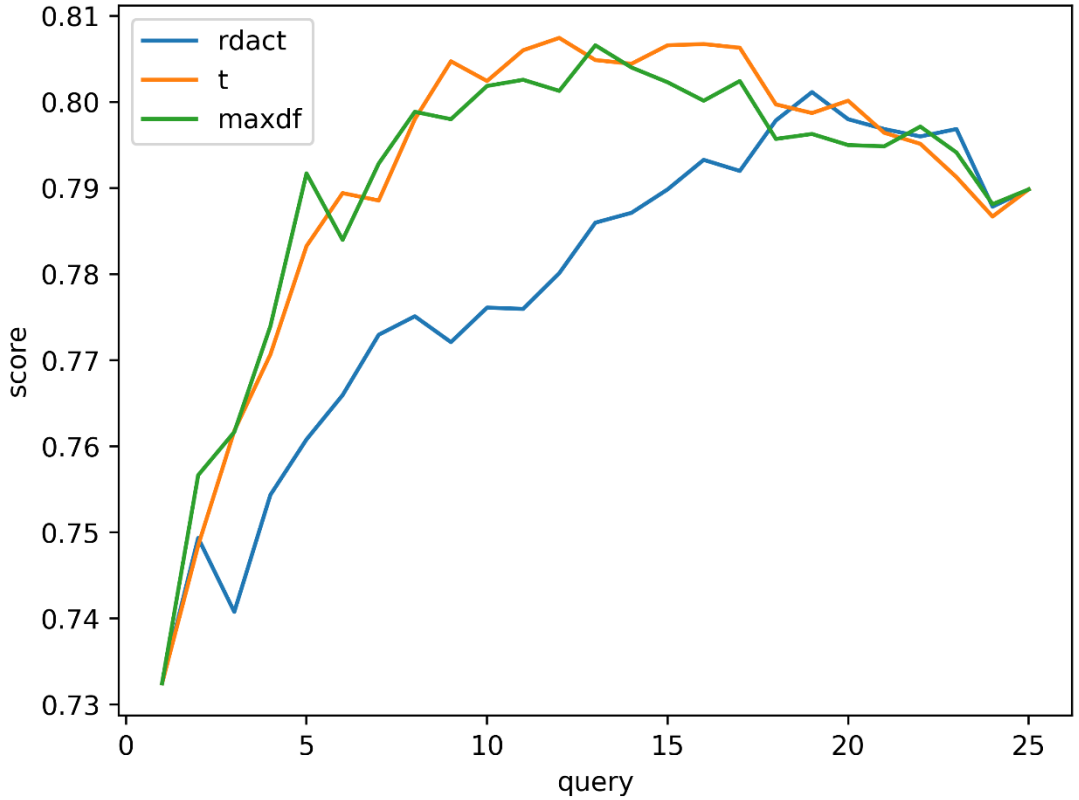

rdact
pal
t
0.76
0.74
0.72
0.70
score
2
4
6
8
10
12
14
16
query

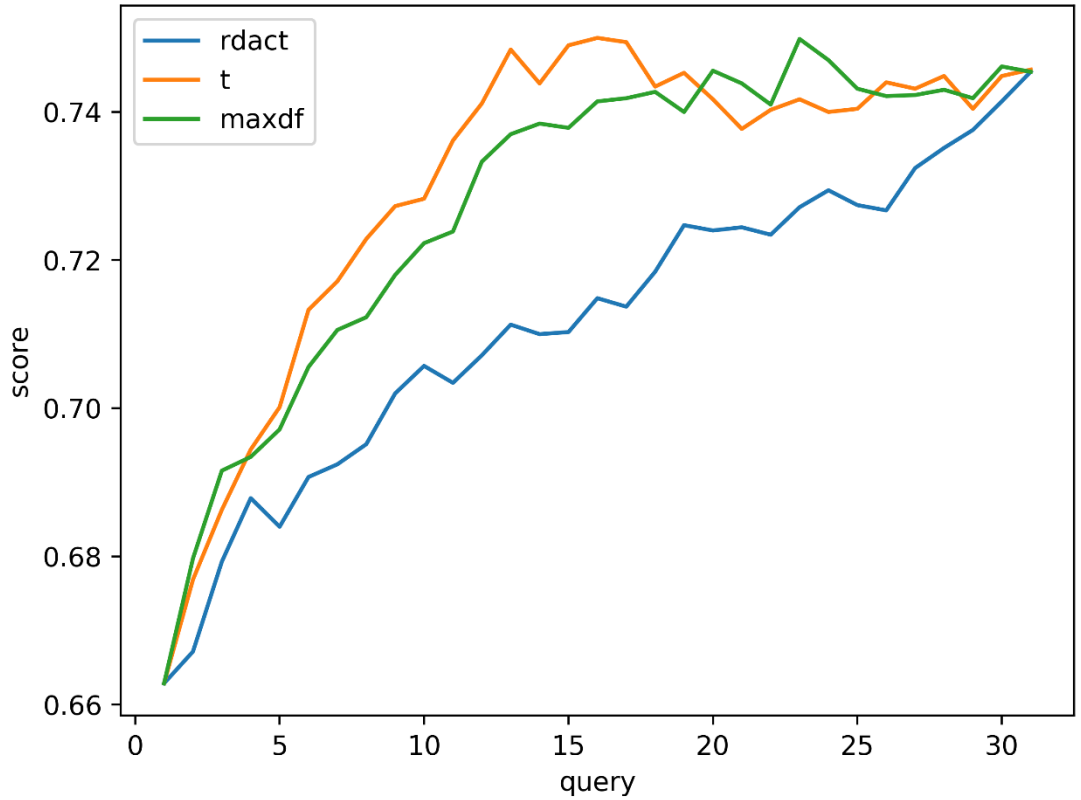

rdact
pal
t
0.80
0.78
0.76
0.74
0.72
score
2
4
6
8
10
12
14
16
query

后验

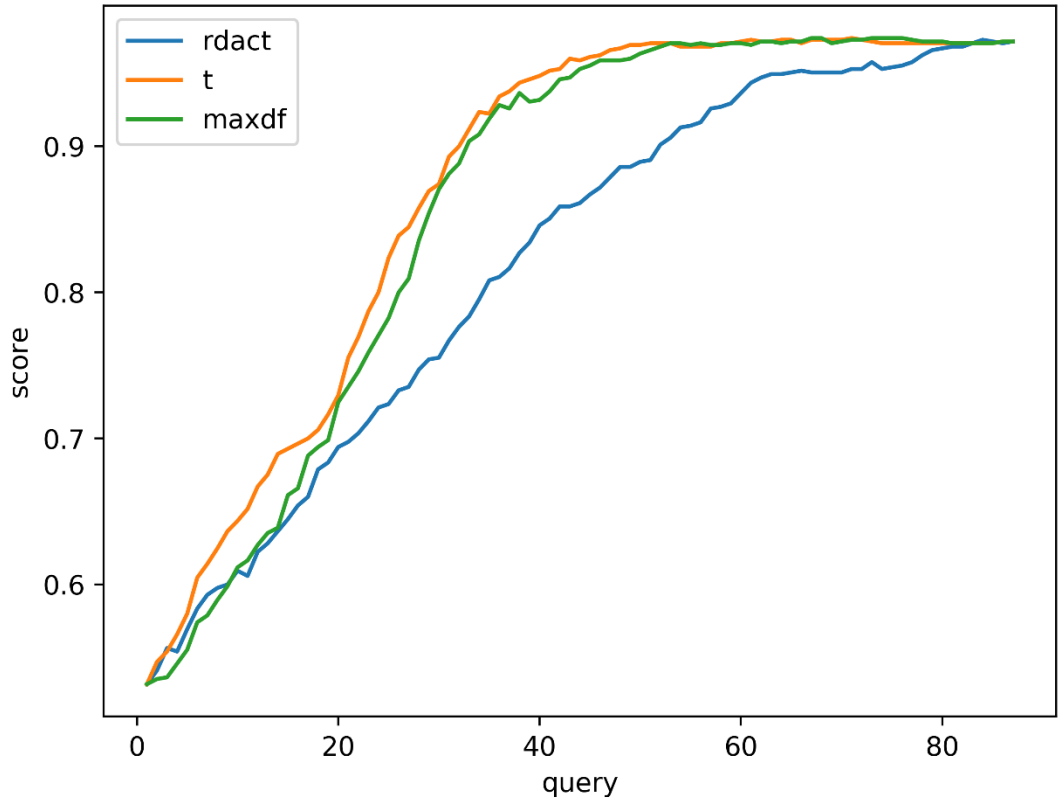

rdact
pal
t
0.82
0.81
0.80
0.79
0.78
0.77
0.76
score
2
4
6
8
10
12
14
16
query

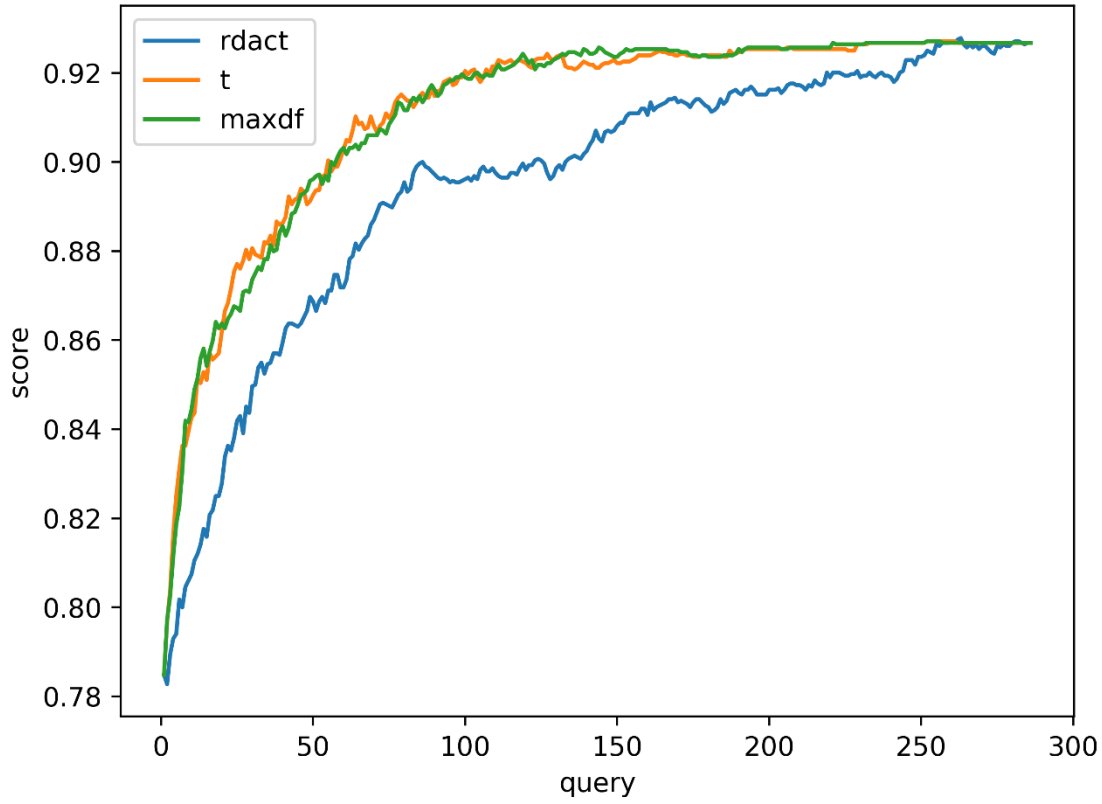
后验+uncertainty



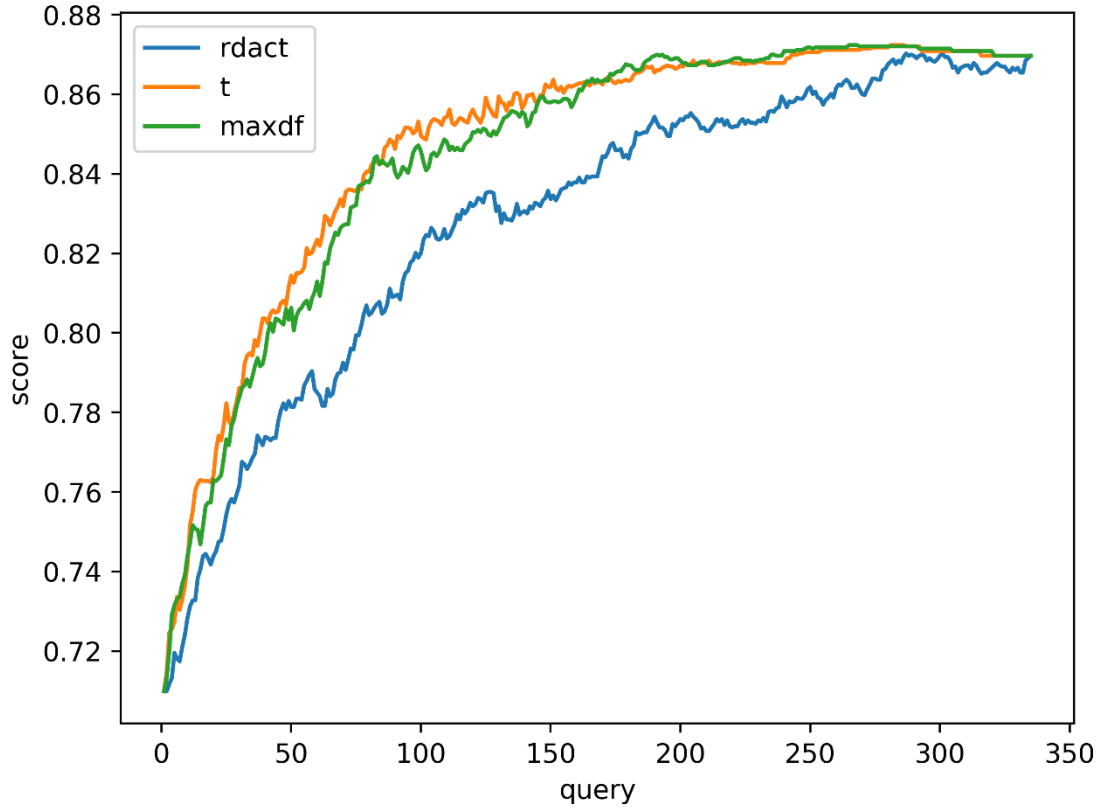


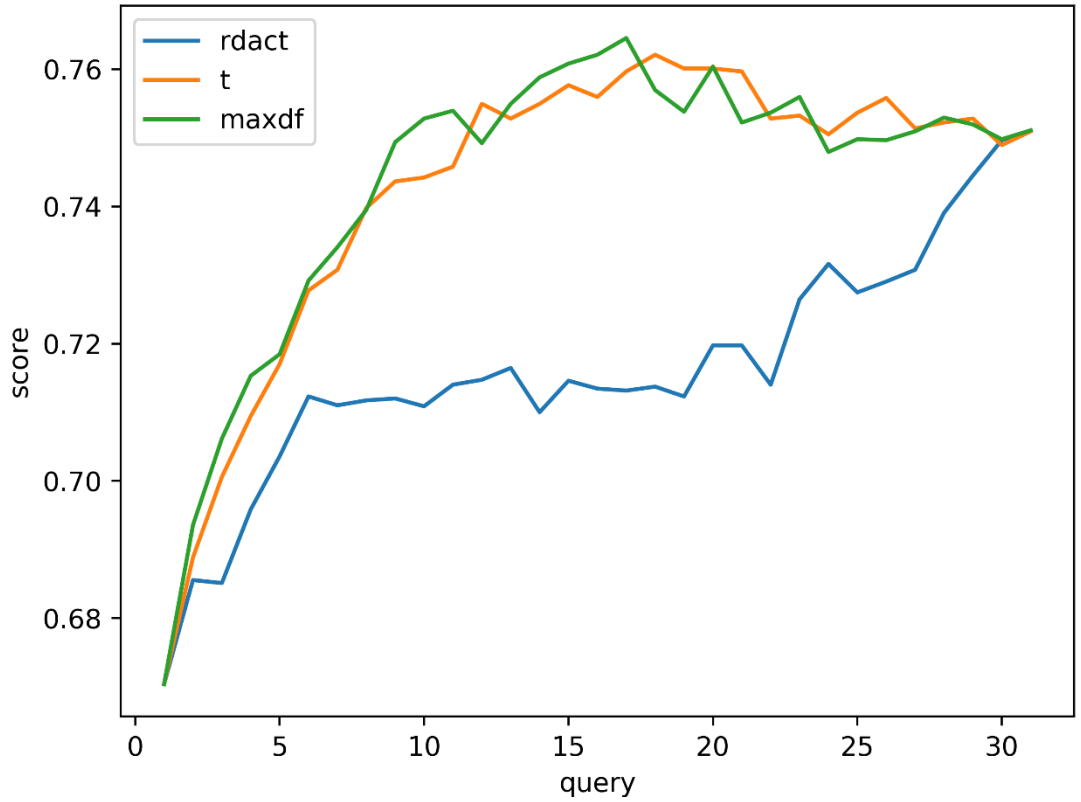
后验+uncertainty



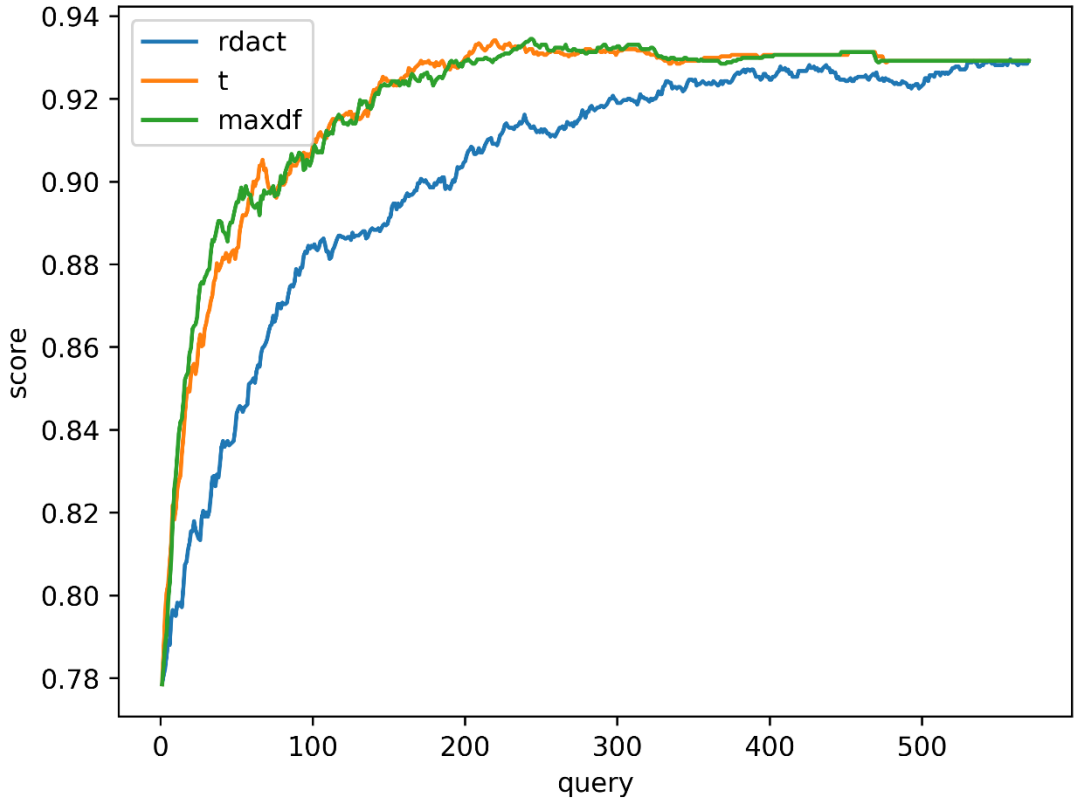


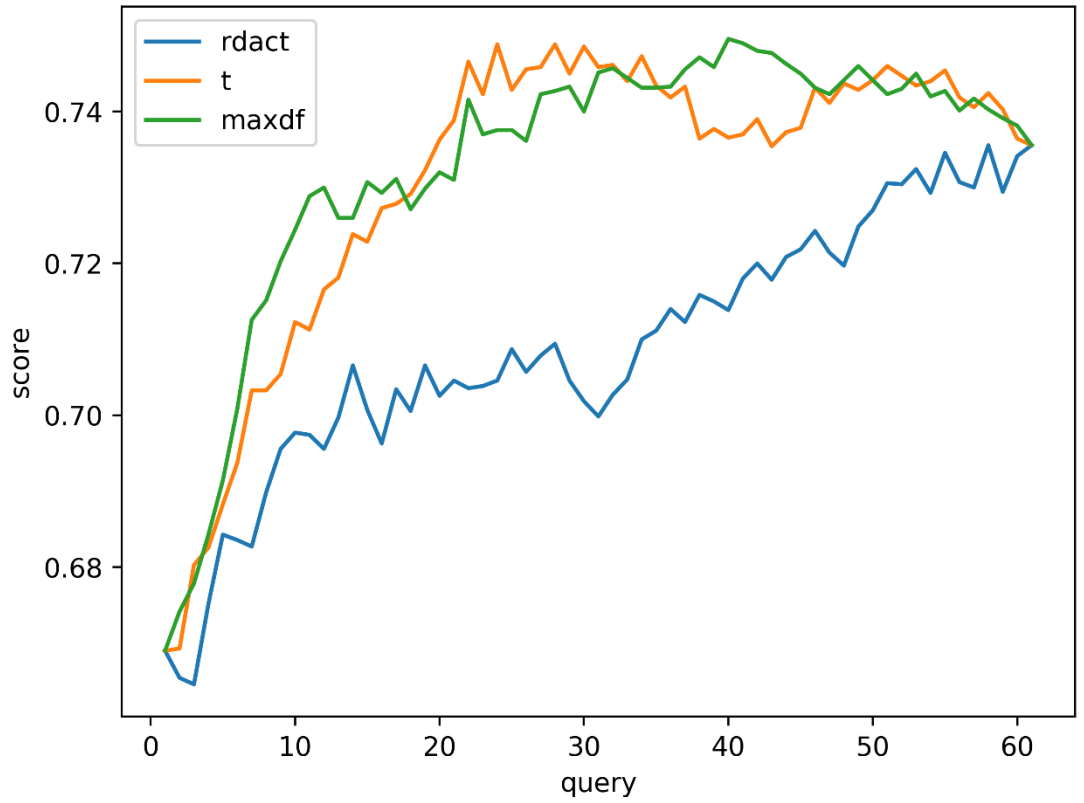
后验+uncertainty



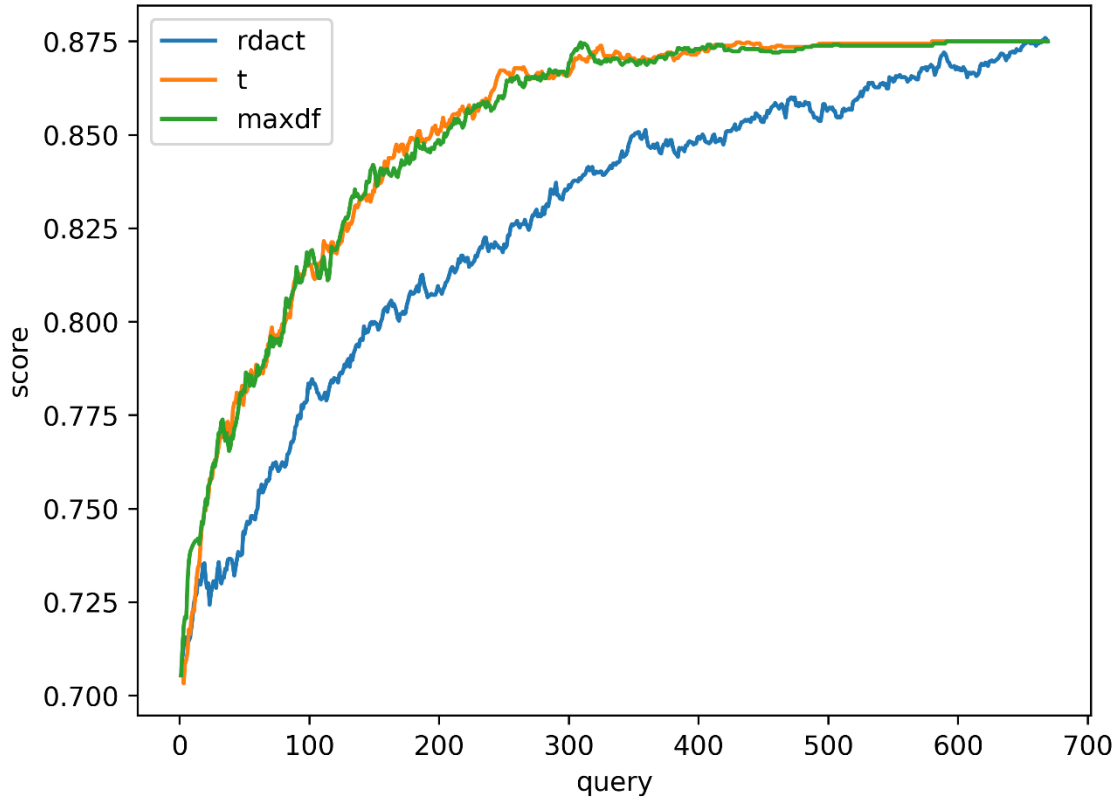


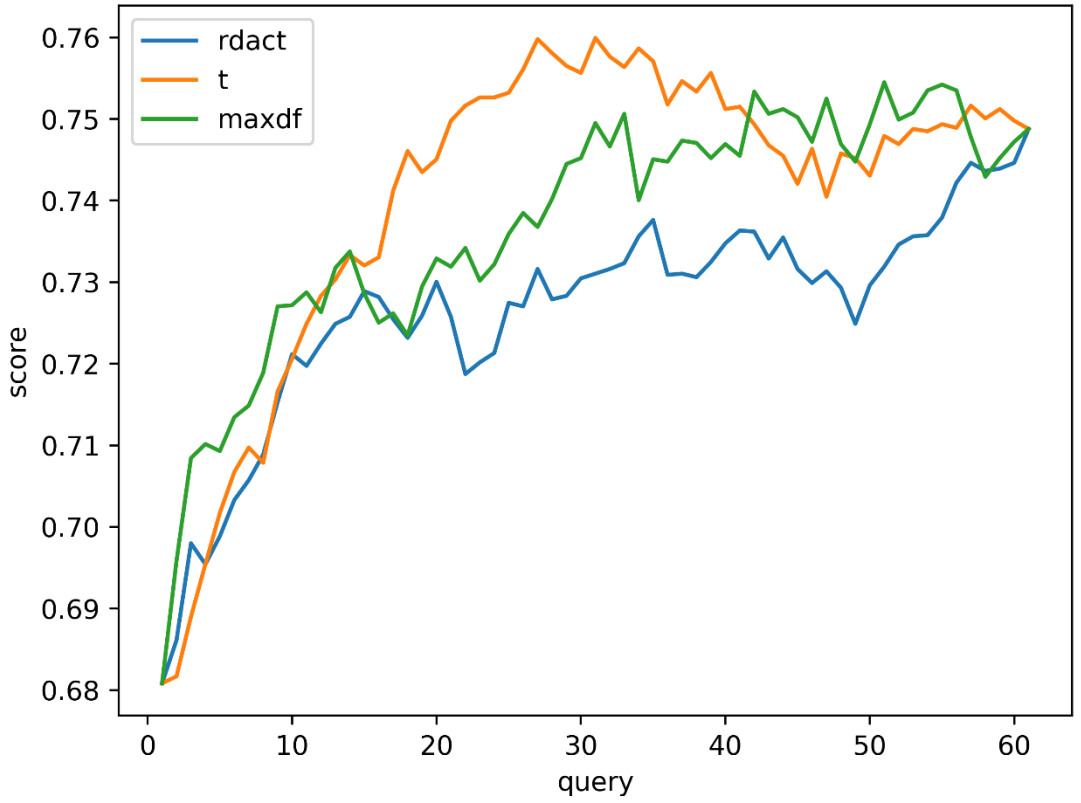
迁移前后差异+uncertainty



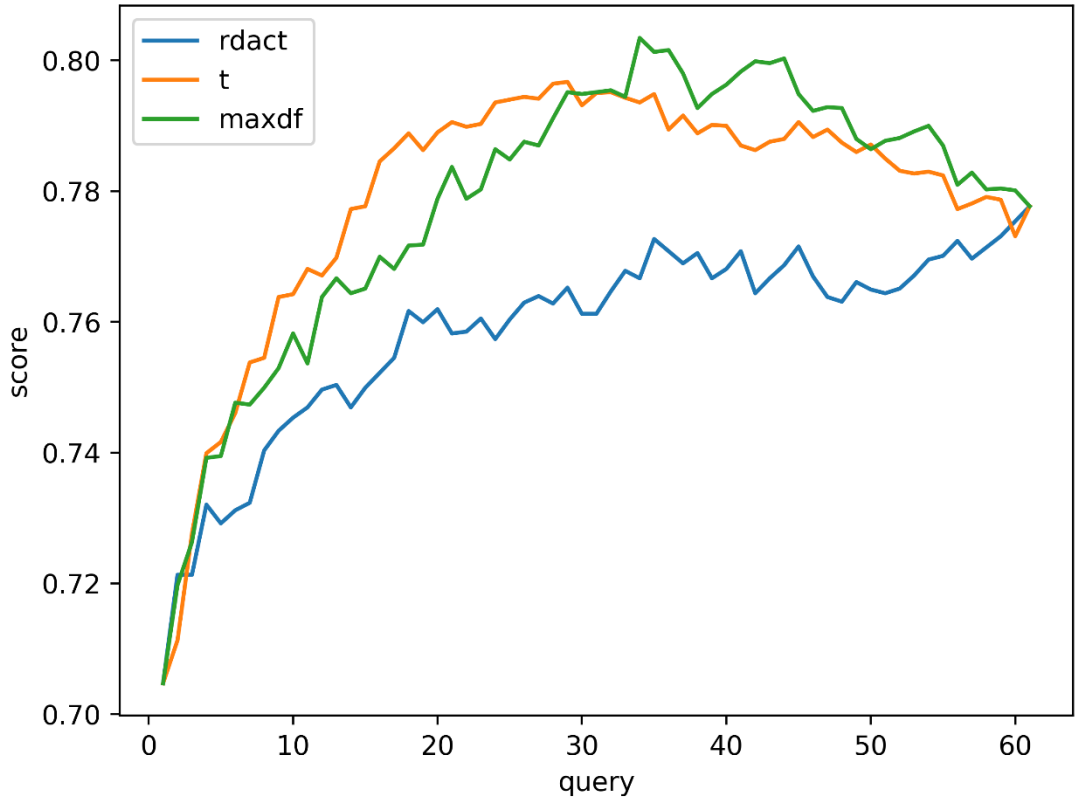


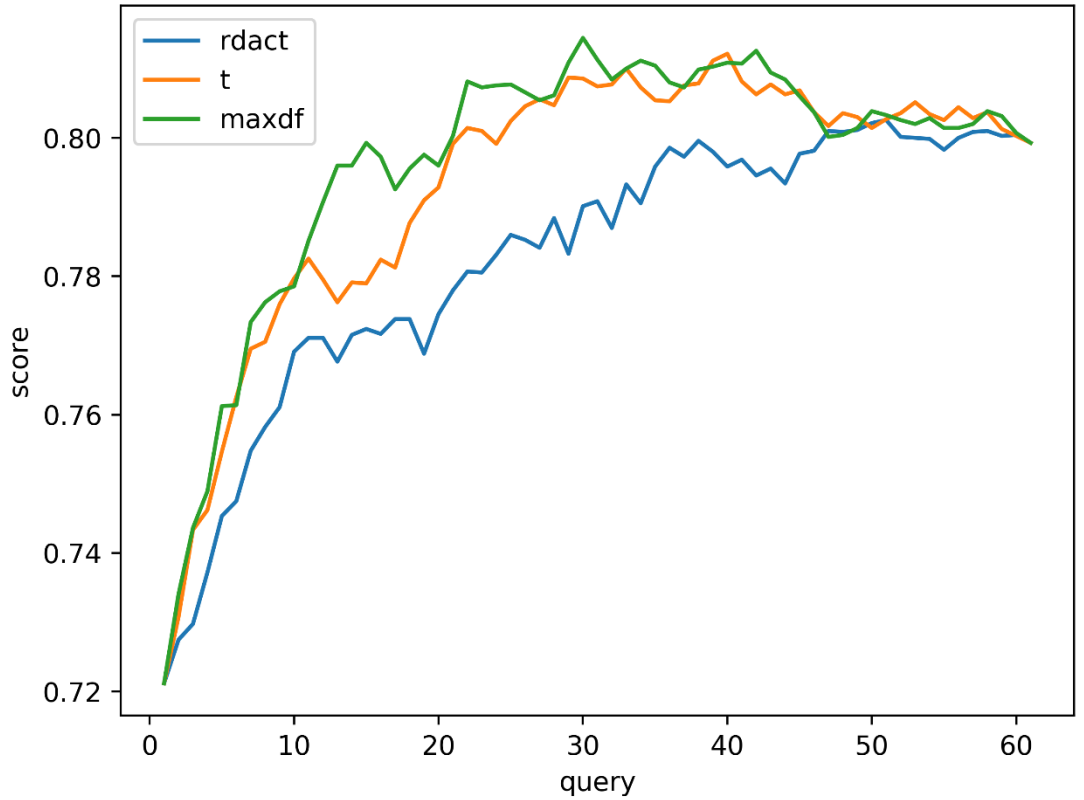
迁移前后差异+uncertainty



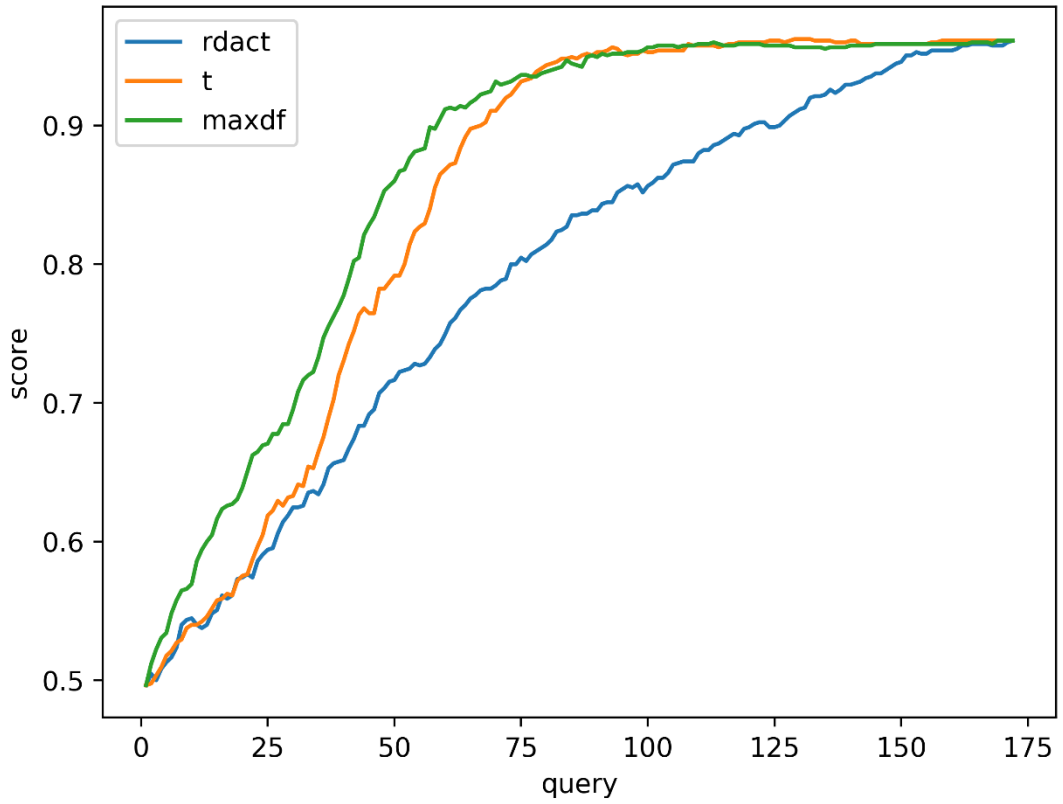


迁移前后差异+uncertainty

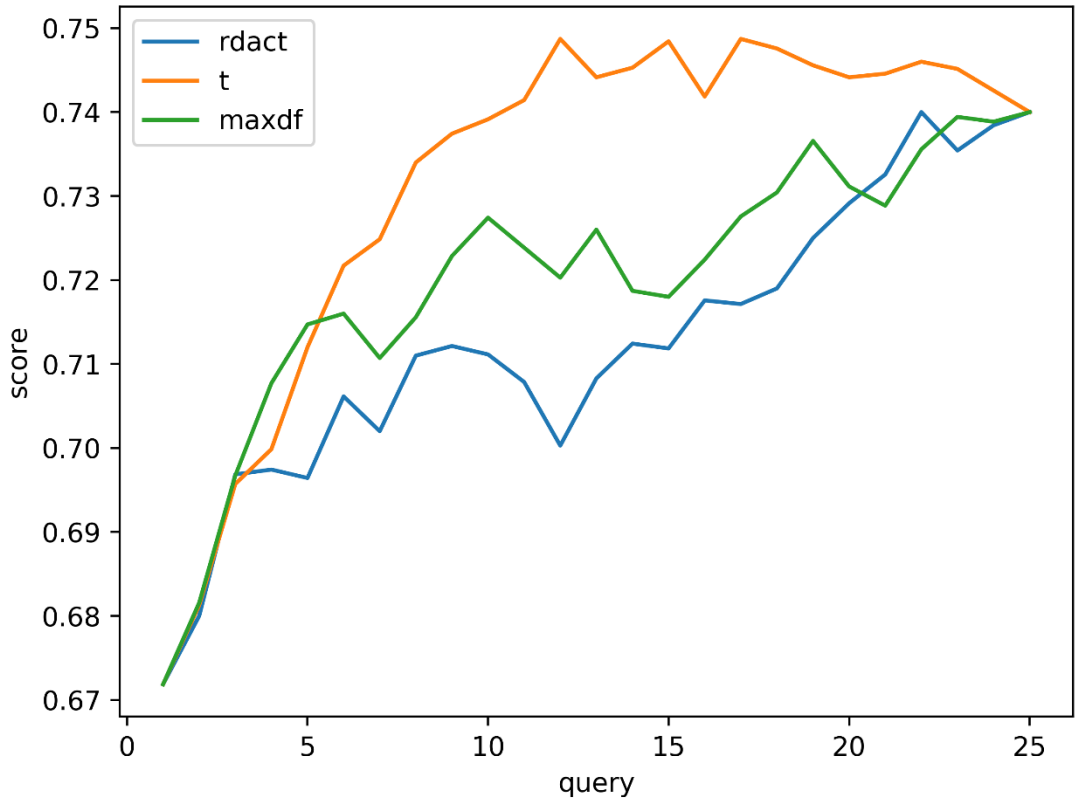


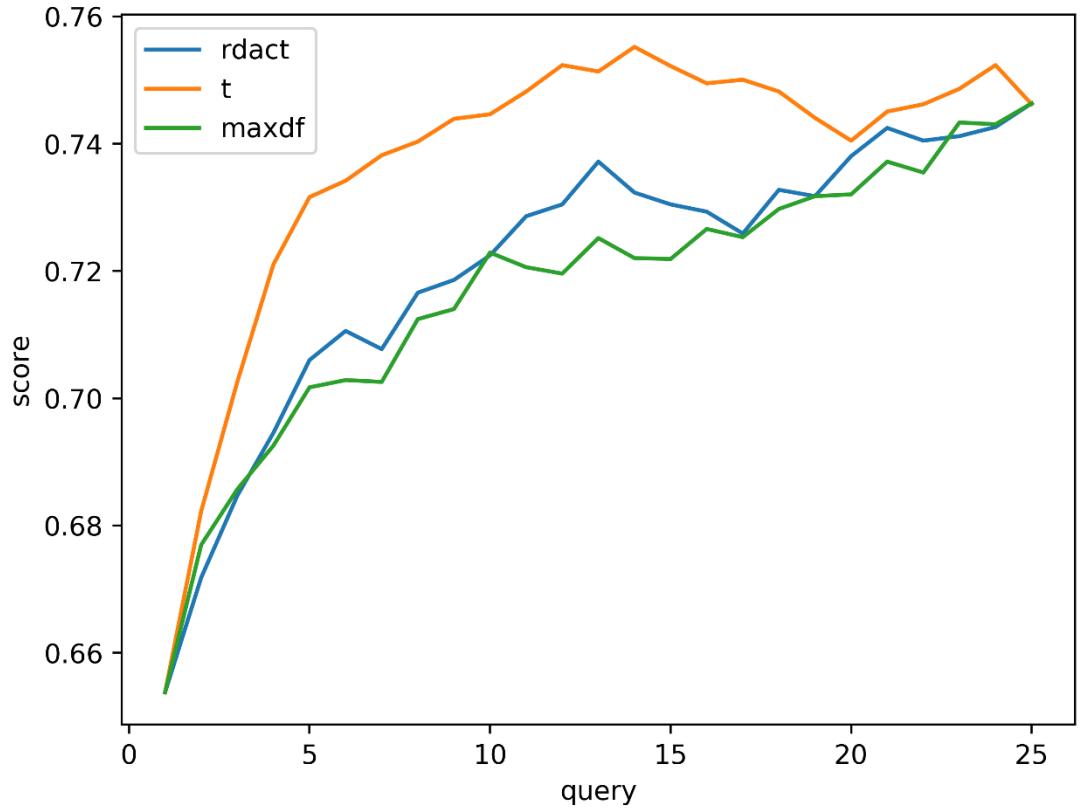


迁移前后差异+uncertainty

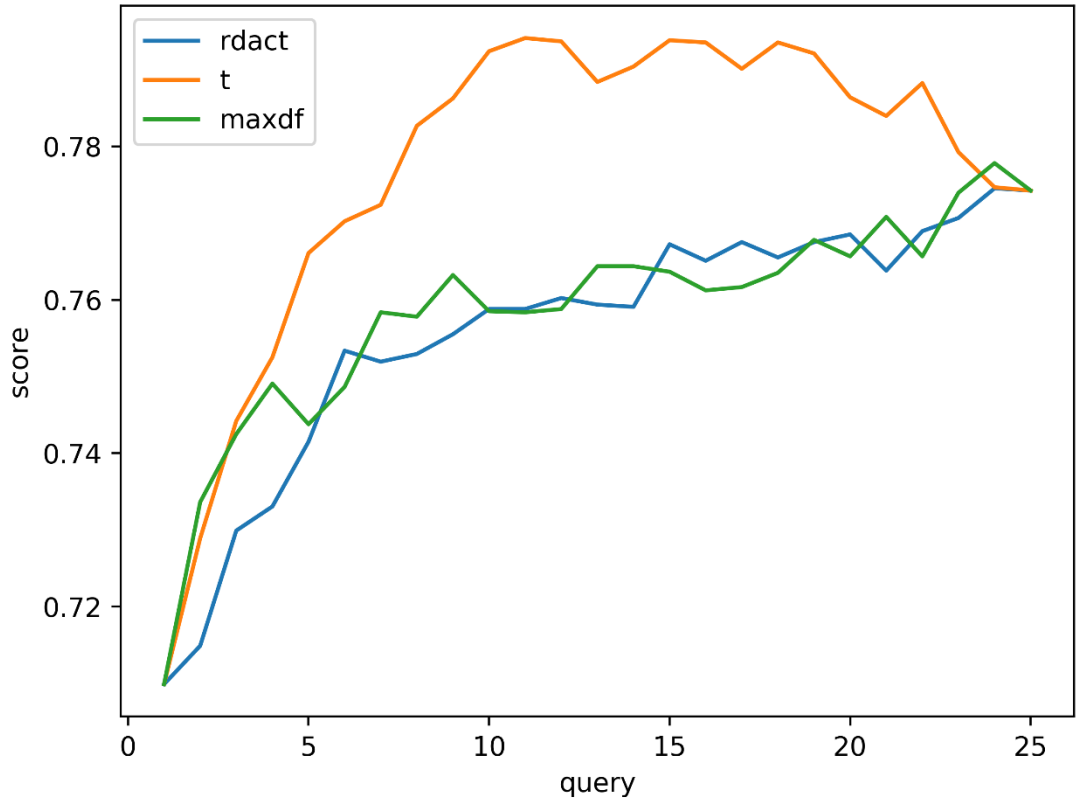


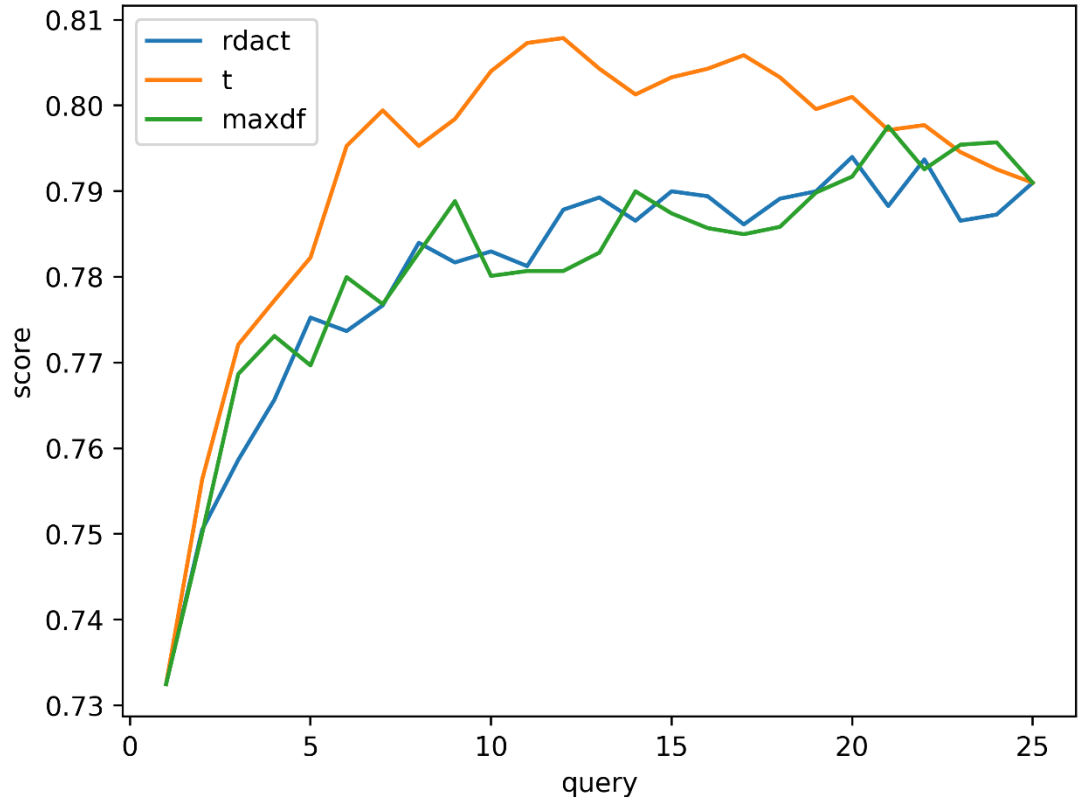
迁移前后差异+后验


rdact
t
maxdf
score
query
0.75
0.74
0.73
0.72
0.71
0.70
0.69
0.68
0.67
0
5
10
15
20
25


rdact
t
maxdf
score
query
0.76
0.74
0.72
0.70
0.68
0.66
0
5
10
15
20
25

迁移前后差异+后验


rdact
t
maxdf
0.78
0.76
0.74
0.72
score
0
5
10
15
20
25
query


rdact
t
maxdf
0.81
0.80
0.79
0.78
0.77
0.76
0.75
0.74
0.73
score
0
5
10
15
20
25
query