

عنوان ارائه:

یک تکنیک برخط مقیاس پذیر و دقیق برای انتخاب ویژگی
در کلان داده‌ها

**Towards Scalable and Accurate Online Feature Selection
For Big Data**

توسط: علیرضا صادقی نسب

استاد: دکتر حسین غفاریان

تاریخ ارائه: ۱۳۹۹/۹/۵

فهرست مطالب

- مقدمه
- معرفی روش
- ارزیابی روش
- بررسی نقاط قوت و ضعف

مقدمه

■ انتخاب ویژگی

☑ با نام‌هایی مانند انتخاب متغیرها، انتخاب مشخصه و انتخاب زیرمجموعه متغیرها نیز شناخته می‌شود

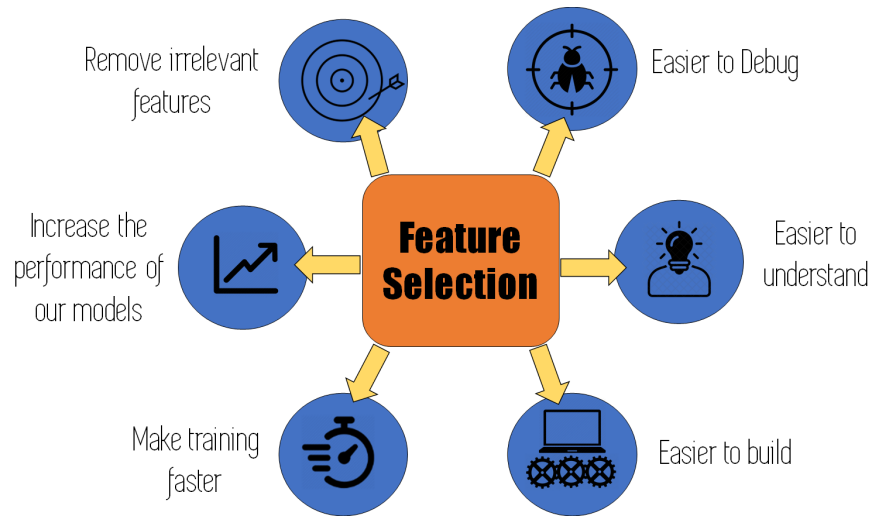
☑ هدف انتخاب ویژگی، انتخاب یک زیرمجموعه از ویژگی‌های به هم مرتبط برای ساخت مدل است

★ دلایل استفاده:

★ ساده‌سازی مدل

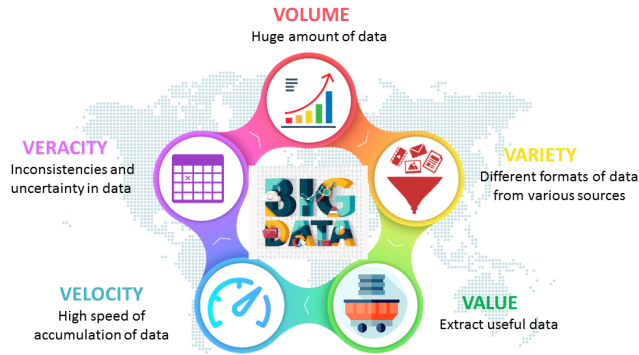
★ کاهش زمان یادگیری

★ پرهیز از نفرین ابعادها



مقدمه

- کلان داده؛ مشخصات، کاربردها و چالش‌ها
- مشخصات:



★ حجم بالا

★ تنوع و گستردگی منابع

★ ارزش نهفته

★ صحت داده

★ سرعت زیاد

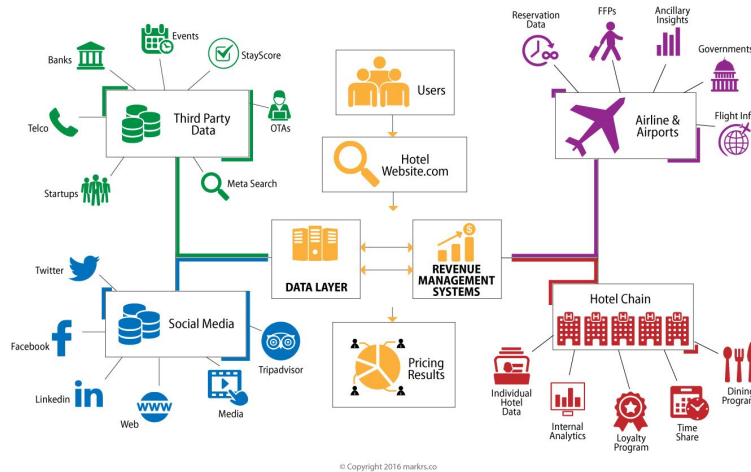
■ کاربردها:

★ بازارهای مالی، بانک‌ها و غیره

★ حمل و نقل

★ بهداشت و درمان

★ اینترنت اشیا



مقدمه

■ چالش‌های اصلی انتخاب ویژگی در کلان داده‌ها

ابعد داده بسیار بالا است 🌐

✓ در مقیاس میلیون بوده و روند رشد بسیار زیاد است

نیاز به مقیاس پذیر بودن 🌐

✓ از روش‌های برخط می‌بایست استفاده شود

✓ هر ویژگی در یک اسکن متوالی پردازش می‌شود

All Features



Feature Selection



Final Features



مقدمه

■ چرا روش‌های دسته‌ای پاسخگو نیستند؟

☑ مقیاس‌پذیری روش‌های انتخاب ویژگی هنگام مواجهه با میلیون‌ها ویژگی، بسیار حیاتی است

☑ با توجه به نتایج به دست آمده از برنامه‌های واقعی، انتظار برای پردازش مجموعه کامل ویژگی‌ها، غیرممکن است

☑ در روش‌های دسته‌ای، تمامی مجموعه ویژگی داده‌های یادگیری باید قابل دسترسی باشد و بعد از آن یک جستجوی سراسری انجام می‌شود



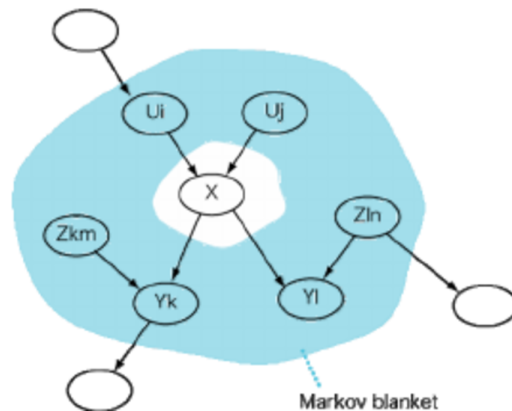
مقدمه

■ پتوی مارکوف

★ در علم آمار و یادگیری ماشین، وقتی کسی بخواهد یک متغیر تصادفی را با مجموعه‌ای از متغیرها استنباط کند، معمولاً یک زیرمجموعه کافی است. به این زیرمجموعه‌ای که شامل تمام اطلاعات مفید باشد، پتوی مارکوف گفته می‌شود

★ اگر یک پتوی مارکوف حداقلی باشد، به این معنی است که هیچ متغیری را بدون از دست دادن اطلاعات رها کند، آن را مرز مارکوف می‌نامند

★ شناسایی پتوی مارکوف یا مرز مارکوف به استخراج ویژگی‌های مفید کمک می‌کند



$$\forall Y \in F - (M \cup \{F_i\}) \text{ s.t. } P(F_i|M, Y) = P(F_i|M).$$

مقدمه

■ متغیر تکراری یا زائد

* یک ویژگی دلخواه $F_i \in F$ اضافی است اگر یک پتوی مارکوف در F داشته باشد

* ویژگی تکراری باید از مجموعه اصلی حذف گردد

* یک روش انتخاب ویژگی، زمانی مطلوب تلقی می‌شود که بتواند ویژگی‌های قویاً مرتبط و غیرتکراری را از مجموعه ورودی، استخراج کند

■ مزایای پردازش مجموعه داده براساس اسکن متوالی

* در هر لحظه، یک بُعد مورد پردازش قرار می‌گیرد

* برای پردازش داده‌های ابعاد بالا، به مشکل محدودیت حافظه برنخواهیم خورد

* برای پردازش، نیازی به مجموعه کامل ویژگی‌های نخواهیم داشت

معرفی روش

■ تعریف مسئله

★ اگر یک مجموعه داده یادگیری $D = \{(d_i, c_i), 1 \leq i \leq N\}$ باشد که N در آن

تعداد نمونه‌ها، هر نمونه داده d_i یک بردار چند بعدی شامل P ویژگی و C یک رده ویژگی شامل K برچسب رده مجزا است

★ مجموعه داده D را نیز می‌توان به صورت $D = \{(F_i, C), 1 \leq i \leq P\}$ بیان کرد.

در اینجا $F_i = \{f_1, f_2, \dots, f_N\}^T$ نشانگر i^{th} ویژگی شامل N نمونه داده و C شامل N برچسب رده می‌باشد

★ در پردازش اسکن متوالی، اگر فرض کنیم F_i ، ویژگی جدید رسیده شده در زمان t_i باشد، می‌توان مسئله را به صورت زیر فرموله کرد:

$$S_{t_i}^* = \arg \min_{S'} \{ |S'| : S' = \arg \max_{\zeta \in \{S \cup F_i\}} P(C | \zeta) \}$$

معرفی روش

■ می توان فرمول قبل را به چند مرحله کلیدی افراز کرد:

☑ تعیین ارتباط F_i و C با استفاده از احتمال شرطی: اگر رابطه پایین صادق باشد، F_i

یک ویژگی غیرمرتبط خواهد بود: $P(C|F_i) = P(C)$

☑ بررسی می کنیم آیا F_i به همراه خود، اطلاعات پیشگویانه اضافی دارد یا خیر. این شرط

به وسیله رابطه روبه‌رو بررسی می شود: $P(C|S_{t_{i-1}}^*, F_i) = P(C|S_{t_{i-1}}^*)$

★ اگر F_i غیرمرتبط یا فاقد اطلاعات باشد، حذف خواهد شد. در غیر این صورت، به $S_{t_{i-1}}^*$

اضافه خواهد شد و $S_{t_i}^*$ در زمان t_i ساخته می شود.

★ در نهایت $S_{t_i}^*$ هرس می شود تا معادله اصلی در آن صادق باشد:

$$S_{t_i}^* = \arg \max_{\zeta \subset S_{t_i}} P(C|\zeta)$$

معرفی روش

📍 معادله‌های دوم و سوم اسلاید قبل در جهت کاهش ویژگی انجام می‌شوند و می‌توان از تعریف متغیر تکراری (یا زائد) استفاده کرد اما از لحاظ پیچیدگی محاسباتی، بسیار سربار دارند زیرا تمام زیرمجموعه‌های $S_{t_{i-1}}^*$ می‌بایست بررسی شوند یعنی $2^{|S_{t_{i-1}}^*|}$ که با توجه به تعداد ابعاد بسیار بالا، عملاً غیرقابل استفاده خواهد بود

★ راه حل چیست؟ به جای محاسبه همبستگی بین ویژگی‌های مشروط بر روی تمام زیرمجموعه‌های ویژگی، از مقایسه‌های جفتی برای محاسبه برخط همبستگی بین ویژگی‌ها استفاده شود

📍 اطلاعات متقابل بین Y و Z به صورت زیر تعریف می‌شود:

$$I(Y; Z) = H(Y) - H(Y | Z)$$

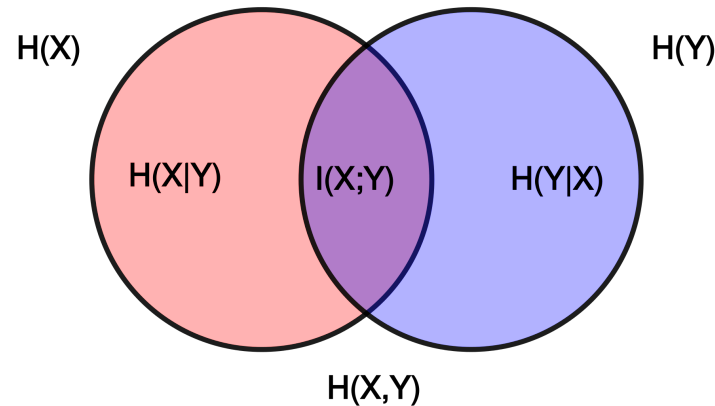
معرفی روش

● آنتروپی ویژگی Y به صورت زیر تعریف می شود:

$$H(Y) = - \sum_{y_i \in Y} P(y_i) \log P(y_i)$$

● آنتروپی ویژگی Y بعد از حضور ویژگی Z به صورت زیر تعریف می شود:

$$H(Y|Z) = - \sum_{z_i \in Z} P(z_i) \sum_{y_i \in Y} P(y_i | z_i) \log P(y_i | z_i)$$



معرفی روش

📍 اگر یک حد آستانه ارتباط مانند δ_1 داشته باشیم و $I(F_i, C) > \delta_1$ باشد، آنگاه می‌گوییم

ویژگی F_i یک ویژگی مرتبط است وگرنه غیرمرتبط بوده و می‌بایست حذف شود

📍 به ازای هر ویژگی Y که عضوی از مجموعه $S_{t_{i-1}}^*$ می‌باشد، اگر $I(F_i; C | Y) = 0$ باشد،

آنگاه می‌گوییم که افزوده شدن ویژگی F_i ، توانایی پیشگویی $S_{t_{i-1}}^*$ را افزایش نخواهد داد

✅ برای برقراری شرط بالا، کافی است دو شرط زیر را بررسی کنیم:


$$I(Y; C) > I(F_i; C) \text{ and } I(F_i; Y) \geq I(F_i; C)$$

★ حال با فرض اضافه شدن ویژگی F_i ، به دنبال ویژگی‌هایی خواهیم بود که پس از درج F_i

قابل حذف هستند. اگر $I(C; Y | F_i) = 0$ آنگاه ویژگی Y حذف خواهد شد

✅ برای برقراری شرط بالا: $I(F_i; C) > I(Y; C) \text{ and } I(Y; F_i) \geq I(Y; C)$

معرفی روش

شبه کد الگوریتم پیشنهادی 

Algorithm 1: The SAOLA Algorithm

Data:

F_i : predictive features; C : the class attribute;


δ_1 : a relevance threshold ($0 \leq \delta_1 \leq 1$);

δ_2 : a correlation bound of $I(F_i; Y)$;

$S_{t_{i-1}}^*$: the selected feature set at time t_{i-1} ;

$S_{t_i}^*$: the selected feature set at time t_i

```
1 repeat
2   Get a new feature  $F_i$  at time  $t_i$ ;
3   /*Solve Eq.(2)*/
4   if  $I(F_i, C) < \delta_1$  then
5     Discard  $F_i$ , and go to Step 18;
6   end
7   for each feature  $Y \in S_{t_{i-1}}^*$  do
8     /*Solve Eq.(3)*/
9     if  $I(Y; C) > I(F_i; C) \ \& \ I(F_i; Y) \geq \delta_2$  then
10      Discard  $F_i$ , go to Step 18;
11    end
12    /*Solve Eq.(4)*/
13    if  $I(F_i; C) > I(Y; C) \ \& \ I(F_i; Y) \geq \delta_2$  then
14       $S_{t_{i-1}}^* = S_{t_{i-1}}^* - Y$ ;
15    end
16  end
17   $S_{t_i}^* = S_{t_{i-1}}^* \cup F_i$ ;
18 until no features are available;
19 Output  $S_{t_i}^*$ ;
```

عمده سربار الگوریتم به جهت محاسبه همبستگی 

بین ویژگی‌ها (مراحل ۴ و ۹ و ۱۳) است

پیچیدگی زمانی الگوریتم به صورت 

$O(P | S_{t_i}^* |)$ یا همان $O(|S_{t_i}^* |)$ است که یک

پیچیدگی خطی است و در مقابل الگوریتم‌های

دیگر، عملکرد بهتری دارد

ارزیابی روش

✓ برای ارزیابی از ۱۴ مجموعه داده که ابعاد

بالایی دارند استفاده شده است

✓ برای مقایسه با روش های برخط انتخاب ویژگی

دیگر، از سه الگوریتم زیر استفاده شده است:

Fast – OSFS ★

Alpha – investing ★

OFS ★

Dataset	# features	#training instances	#testing instances
madelon	500	2,000	600
hiva	1,617	3,845	384
leukemia	7,129	48	24
lung-cancer	12,533	121	60
ohsumed	14,373	3,400	1,600
breast-cancer	17,816	190	96
dexter	20,000	300	300
apcj-etiology	28,228	11,000	4,779
dorothea	100,000	800	300
thrombin	139,351	2,000	543
news20	1,355,191	9,996	10,000
url1	3,231,961	20,000	20,000
webspam	16,609,143	20,000	78,000
kdd2010	29,890,095	20,000	100,000

✓ برای مقایسه با روش های دسته ای انتخاب ویژگی

دیگر، از سه الگوریتم زیر استفاده شده است:

GDM | SPSF – LAR | FCBF ★

ارزیابی روش

مقایسه با روش‌های برخط 

Dataset	SAOLA	Fast-OSFS	Alpha-investing
dexter	0.7600	0.7800	0.5000
lung-cancer	0.9833	0.9667	0.9167
hiva	0.9635	0.9635	0.9531
breast-cancer	0.6771	0.6667	0.5833
leukemia	0.9167	0.7917	0.6250
madelon	0.5617	0.5283	0.5767
ohsumed	0.9275	0.9306	0.9325
apcj-etiology	0.9793	0.9702	0.9851
dorothea	0.9613	0.9457	0.7400
thrombin	0.9374	0.9300	0.9371
average rank w/t/l	2.45 -	1.85 5/4/1	1.70 6/3/1

Dataset	SAOLA	Fast-OSFS	Alpha-investing
dexter	0.8133	0.8200	0.5000
lung-cancer	0.9500	0.9000	0.8333
hiva	0.9661	0.9635	0.9635
breast-cancer	0.6042	0.6771	0.7188
leukemia	0.9583	0.9583	0.6667
madelon	0.6083	0.6100	0.6067
ohsumed	0.9437	0.9450	0.9331
apcj-etiology	0.9872	0.9868	0.9828
dorothea	0.9343	0.9371	0.9343
thrombin	0.9613	0.9595	0.9613
average rank w/t/l	2.25 -	2.30 1/8/1	1.45 4/5/1

میزان دقت با استفاده از الگوریتم طبقه‌بند KNN

میزان دقت با استفاده از الگوریتم طبقه‌بند J48

Dataset	SAOLA	Fast-OSFS	Alpha-investing
dexter	21	9	1
lung-cancer	35	6	7
hiva	12	5	48
breast-cancer	46	7	2
leukemia	17	5	2
madelon	3	3	4
ohsumed	65	11	297
apcj-etiology	75	67	634
dorothea	63	5	113
thrombin	20	9	60

Dataset	SAOLA	Fast-OSFS	Alpha-investing
dexter	3	4	6
lung-cancer	6	4	2
hiva	1	36	7
breast-cancer	5	4	3
leukemia	2	2	1
madelon	0.1	0.1	0.1
ohsumed	6	343	497
apcj-etiology	22	> 3 days	9,781
dorothea	58	375	457
thrombin	63	18,576	291

تعداد ویژگی‌های انتخاب شده

زمان اجرا (برحسب ثانیه)

ارزیابی روش

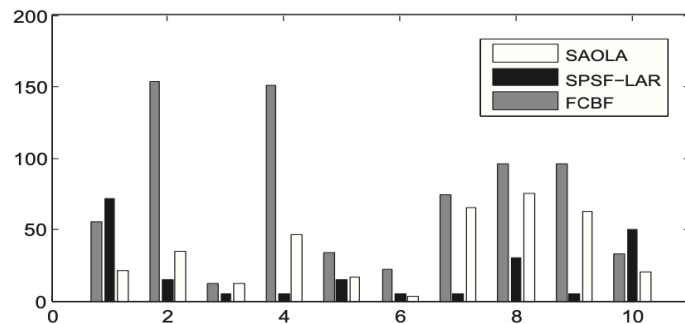
مقایسه با روش‌های دسته‌ای 

Dataset	SAOLA	FCBF	SPSF-LAR
dexter	0.7600	0.7967	0.7233
lung-cancer	0.9833	0.9500	0.9833
hiva	0.9635	0.9609	0.9635
breast-cancer	0.6771	0.6563	0.6771
leukemia	0.9167	1.0000	1.0000
madelon	0.5617	0.5767	0.5633
ohsumed	0.9275	0.9300	0.9113
apcj-etiology	0.9793	0.9826	0.9803
dorothea	0.9613	0.9200	0.8857
thrombin	0.9374	0.9429	0.9650
average rank	2.10	1.85	2.05
w/t/l	-	3/4/3	3/5/2

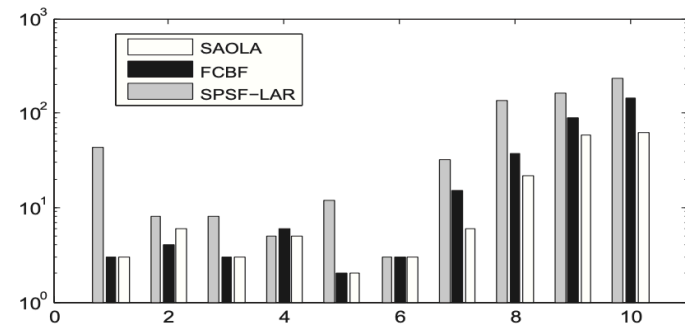
Dataset	SAOLA	FCBF	SPSF-LAR
dexter	0.8133	0.8567	0.8700
lung-cancer	0.9500	0.9500	0.9833
hiva	0.9661	0.9661	0.9635
breast-cancer	0.6042	0.6042	0.6458
leukemia	0.9583	0.9583	0.9583
madelon	0.6083	0.6067	0.6183
ohsumed	0.9437	0.9444	0.9431
apcj-etiology	0.9872	0.9866	0.9872
dorothea	0.9343	0.9314	0.9029
thrombin	0.9613	0.9576	0.9558
average rank	1.85	2.15	2.00
w/t/l	-	0/9/1	1/5/4

میزان دقت با استفاده از الگوریتم طبقه‌بند KNN

میزان دقت با استفاده از الگوریتم طبقه‌بند J48



تعداد ویژگی‌های انتخاب شده



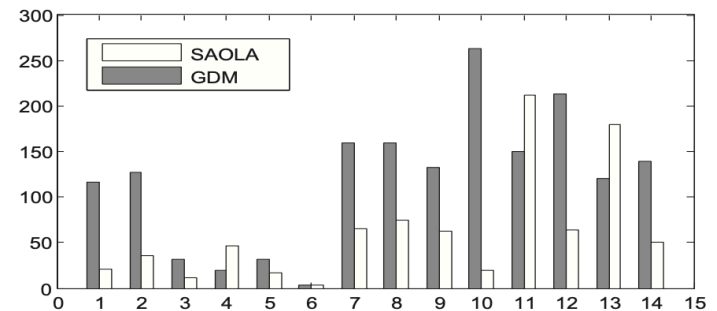
زمان اجرا (برحسب ثانیه)

ارزیابی روش

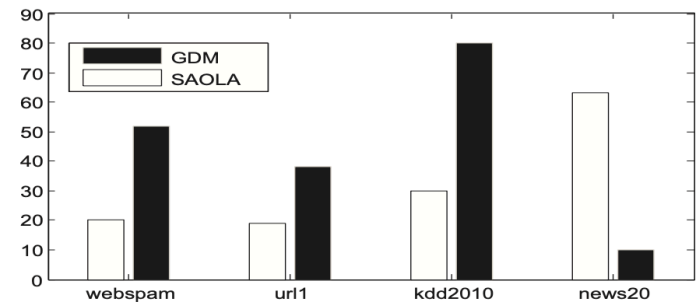
مقایسه با روش دسته‌ای *GDM*

Dataset	KNN		J48	
	SAOLA	GDM	SAOLA	GDM
dexter	0.7600	0.9100	0.8133	0.9100
lung-cancer	0.9833	0.9833	0.9500	0.9833
hiva	0.9635	0.9661	0.9661	0.9661
breast-cancer	0.6771	0.4792	0.6042	0.4792
leukemia	0.9167	1.0000	0.9583	1.0000
madelon	0.5617	0.5833	0.6083	0.5833
ohsumed	0.9275	0.9438	0.9437	0.9438
apcj-etiology	0.9793	0.9879	0.9872	0.9879
dorothea	0.9613	0.9371	0.9343	0.9371
thrombin	0.9374	0.7300	0.9613	0.7300
news20	0.7755	0.7354	0.8276	0.7354
url1	0.9627	0.9765	0.9744	0.9765
kdd10	0.878	0.8179	0.8723	0.8179
webspam	0.9532	0.9617	0.9611	0.9617
Ave rank	1.3929	1.6071	1.3929	1.6071
w/t/l	-	5/4/5	-	5/6/3

میزان دقت با استفاده از الگوریتم‌های طبقه‌بند



تعداد ویژگی‌های انتخاب شده



زمان اجرا (بر حسب ثانیه)

بررسی نقاط قوت و ضعف

☆ نقاط قوت

- ☑ توانا در انتخاب ویژگی در کلان داده‌ها با ابعاد بسیار زیاد
- ☑ بهره‌مندی از مقایسه‌های جفتی جهت کاهش سربارهای محاسباتی
- ☑ دقت بالا همراه سرعت بالا به جهت استفاده از رویه اسکن برخط

☆ نقاط ضعف

- ☑ در هر لحظه، فقط یک بُعد پردازش می‌شود
- ☑ از اطلاعات گروهی پشتیبانی نمی‌کند؛ برای مثال در آنالیز عکس، ویژگی‌های در یک گروه تولید می‌شوند و رنگ، الگو و سایر اطلاعات بصری را نشان می‌دهند
- ☑ فرض شده است که هر نمونه فقط متعلق به یک برچسب است

با تشکر از توجه شما