عنوان ارائه:

ک-وارپ: گمنام‌سازی داده‌های جریانی گوناگون به وسیله پارتیشن‌بندی

K-VARP: K-anonymity for varied data streams via partitioning

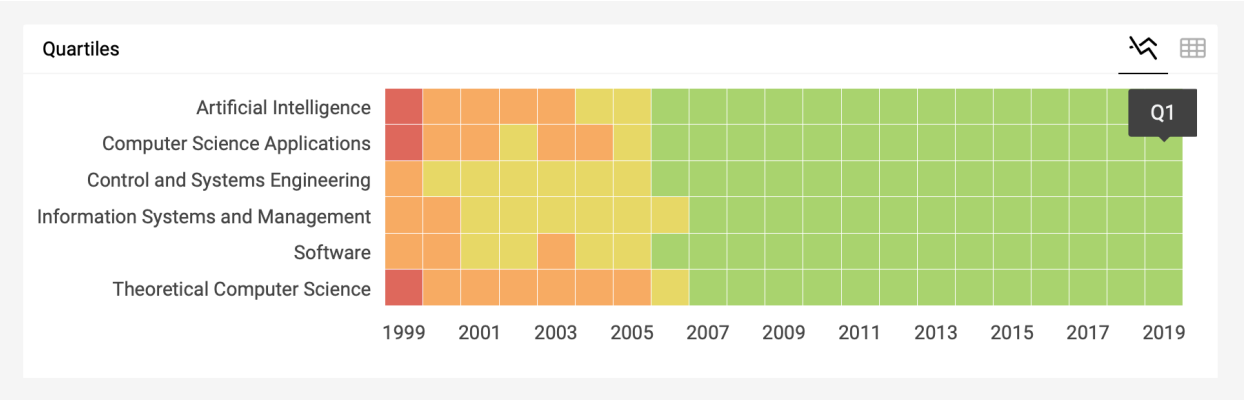
توسط: علیرضا صادقی نسب

استاد: دکتر حسین غفاریان

تاریخ ارائه: ۱۴۰۰/۲/۲۲

مقدمه

- اطلاعات مقاله
  - عنوان: $K-VARP: K-Anonymity\ for\ varied\ data\ streams\ via\ partitioning$
  - سال چاپ: 2018
  - تعداد ارجاع: 15
  - مجله: *Information Sciences*
  - ناشر: *Elsevier*



اطلاعات نویسندگان

Ankhbayar Otgonbayar

University of the west of Scotland
Verified email at uws.ac.uk

Internet Of Things Artificial Intelligence Deep learning Data stream processing Privacy preservation

Citations	70	64
h-index	3	3
i10-index	3	3

Zeeshan PERVEZ

University of the West of Scotland
Verified email at uws.ac.uk - Homepage

Internet-of-Things Cyber Security Secure Cloud Services Data Stream Processing Data Analysis

Citations	939	631
h-index	17	13
i10-index	33	20

Keshav Dahal

Professor in Intelligent Systems at the University of the West of Scotland
Verified email at uws.ac.uk

Intelligent Systems Artificial Intelligence Optimisation Scheduling Operational Research

Citations	3294	1652
h-index	31	19
i10-index	74	45

فهرست مطالب

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

مقدمه

- کلان‌داده و اهمیت آن

☆ امروزه تمامی صنعت‌ها و پژوهش‌ها توسط کلان‌داده‌ها تغذیه می‌شوند

☆ از جمله کاربردهای کلان‌داده:

☆ داده‌های پزشکی بیمارستان‌ها و مراکز درمانی

☆ داده‌های موسسات مالی، بورس و ...

☆ داده‌های کاربران در شبکه اجتماعی

☆ ویژگی کلان‌داده‌ها را اصطلاحاً با $8V$ توصیف می‌کنند

مقدمه

- رایانش ابری

☆ با ظهور کلان‌داده‌ها و بروز مشکلاتی در مدیریت آن‌ها، پردازش ابری بسیار پرکاربرد شد

☆ به طور کلی به در دسترس بودن منابع و قدرت محاسباتی بدون مدیریت مستقیم کاربر تاکید دارد

☆ الگویی تازه برای عرضه، مصرف و تحویل خدمات رایانشی با به کارگیری شبکه است

☆ مزایای رایانش ابری:

☑ دسترسی مقیاس‌پذیر پویا به مزایای تکنولوژی بدون داشتن دغدغه از استقرار، نگه‌داری و عملیات زیرساخت فیزیکی

☑ ارایه خدمات به صورت بستر یا زیرساخت برای استقرار، اجرا و شبیه‌سازی بلادرنگ

مقدمه

- لزوم حفظ حریم خصوصی در سامانه‌های مبتنی بر رایانش ابری

* از آنجایی که در اغلب موارد، داده‌ها از دستگاه‌های کاربران جمع‌آوری می‌شود، شناسایی و دستیابی به آن‌ها توسط یک شخص مخرب، بسیار زیان‌آور خواهد بود
* شخص مخرب می‌تواند به بکارگیری داده‌های دیگر، رفتارهای کاربران را بیاموزد
* محبوب‌ترین راه‌حل برای حفظ حریم خصوصی، گمنام‌سازی است

مقدمه

- گمنام‌سازی در داده‌ها
  - گمنام‌سازی پایگاه‌داده‌ای
    - فرآیند بر روی یک مجموعه‌داده جمع‌آوری شده ثابت انجام می‌شود
    - هدف اصلی، کاهش میزان از دست دادن اطلاعات است
  - گمنام‌سازی داده‌های جریانی
    - در داده‌های جریانی، زمان بسیار مهم است زیرا داده‌ها پس از یک میزان تاخیر، منقضی شده و بلااستفاده می‌شوند
    - فرآیند گمنام‌سازی به صورت پویا انجام می‌شود

Publication delay ⟺ Information loss

مقدمه

- گم شدن داده در جریان داده؛ چرا و چطور
  - دستگاه‌هایی که با سامانه‌های ابری در حال تبادل داده هستند، استاندارد واحدی ندارند. از این رو، ممکن است داده‌هایی که ارسال می‌کنند حاوی نقص باشند
  - اصطلاحاً به این داده‌ها، داده‌های جریانی متنوع گفته می‌شود زیرا در هر داده‌ای که حاوی مقادیر گم شده است، مجموعه شبه‌صفات متفاوت خواهد بود
  - ۳ دلیل اصلی وجود (تولید) داده‌های گم شده:
    - تنظیمات متفاوت در دستگاه‌های کاربران
    - الگوی استفاده متفاوت
    - شرایط محیطی غیرقابل پیش‌بینی

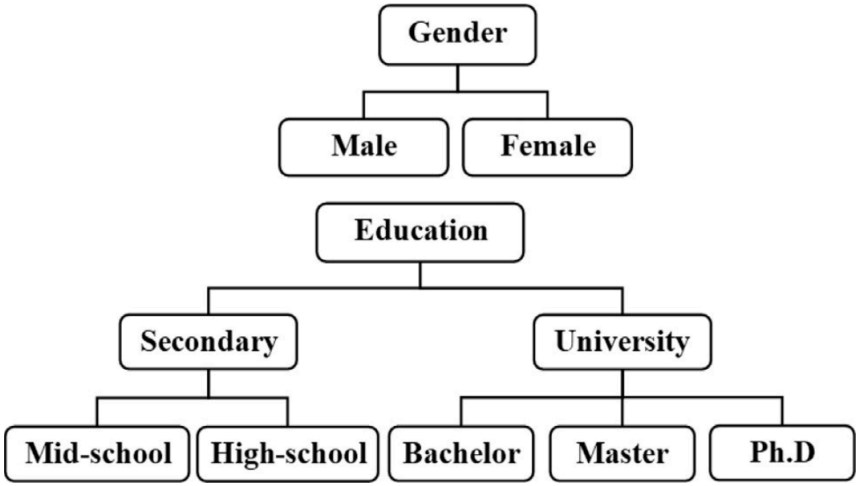
مقدمه

- با داده‌های گم شده چکار کنیم؟

* نسبت‌دادن یا *imputation*: جهت بازسازی داده‌های گم شده، مقادیر مناسب از پیش محاسبه شده جایگذاری مقدار گم شده می‌شود
* حاشیه‌سازی یا *marginalization*: در این حالت، به داده اصلی، چیزی اضافه نمی‌شود. با داده‌های گم شده در این حالت، مانند مقدار *NULL* رفتار می‌شود
* افرازبندی یا *partitioning*: در این حالت، مجموعه‌داده به چند زیرمجموعه بدون مقادیر گم شده تقسیم و تبدیل می‌شود و سپس روش‌های گمنام‌سازی بر روی هر یک از آن‌ها، اعمال می‌شود.

تعاریف پایه

- سلسله‌مراتب تعمیم دامنه



- عمومی‌سازی داده‌ها

	Age	Gender	Education
t_1 ***(Tuple)***	20	Male	Bachelor
G_1 ***(Generalization)***	[20–24]	Gender	University

تعاریف پایه

- اتلاف اطلاعات

$$InfoLoss(t, G_t) = \frac{1}{|G_t|}\left(\sum_{q_i \in Q_t} Loss(v_{qi})\right)$$

$$Loss(v_{qi}) = \begin{cases} \frac{r_{i.u} - r_{i.l}}{R_{i.u} - R_{i.l}} & if\ g_i \in [r_{i.l}, r_{i.u}] \\ \frac{|leaves(H_i)|-1}{|leaves(DGH_i)|-1} & if\ g_i = H \end{cases}$$

$$Loss(v_{Age}) = \frac{|24-20|}{|100-0|} = 0.04$$

$$Loss(v_{Gender}) = \frac{|leaves(Gender)|}{|leaves(Gender)|} = \frac{2}{2} = 1$$

$$Loss(v_{Education}) = \frac{|leaves(University)|}{|leaves(Education)|} = \frac{3}{7} = 0.428$$

میانگین مجموع
میزان اتلاف اطلاعات

$$AverageInfoLoss(N) = \frac{1}{N}\sum_{i=1}^{N} InfoLoss(t_i, G_i)$$

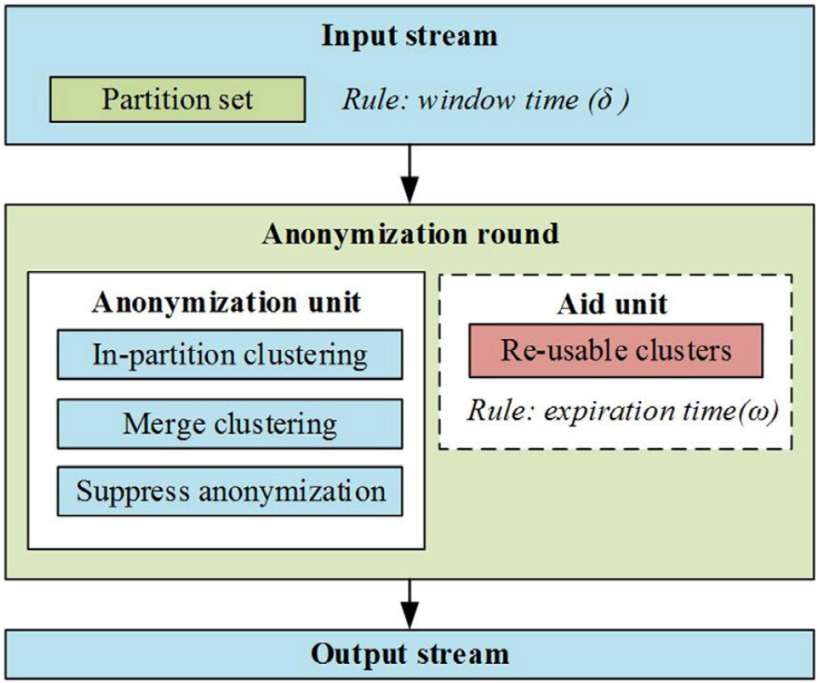
تعاریف پایه

- طبقه‌بندی داده‌های گم‌شده

  - گم‌شده غیرتصادفی($NMAR$): علت گم شدن مقادیر داده به صورت واضح مشخص است و همبستگی مستقیم بین این دو وجود دارد. به عنوان مثال: خالی بودن صندلی‌های هواپیما
  - گم‌شده تصادفی(MAR): تصادفی بودن گم شدن داده‌ها علت دارد ولی خود گم شدگی، علتی ندارد. به عنوان مثال: رد کردن برخی از ایرادات گرامری و نوشتاری توسط ویراستار
  - گم‌شده کاملاً تصادفی($MCAR$): هیچ توضیحی برای علت گم شدن مقادیر داده وجود ندارد و کاملاً براساس اتفاق پیش می‌آید

معرفی روش

- نمای کلی الگوریتم $K - VARP$



معرفی روش

- نمای کلی الگوریتم

☆ از پنجره کشویی مبتنی بر زمان استفاده می‌کند

☆ داده‌هایی که به تازگی به سیستم می‌رسند، براساس مجموعه شبه‌صفاتشان، در پارتیشن‌های S_p قرار می‌گیرند

☆ زمان ورود هر داده نگه‌داری می‌شود زیرا انقضای داده‌ها نیز حائز اهمیت است

Algorithm 1 $K - VARP(VS, K, \delta, \omega, R)$.

```
1: Let S_p be a set of partitions which will be used as a buffer, initialized empty;
2: Let S_k be a set of K-anonymous clusters which will be re-used, initialized empty;
3: while VS ≠ NULL do
4:     Read tuple t_i from VS and assign partition of S_p or create new partition for t_i;
5:     if Oldest tuple in buffer is expiring then
6:         TriggerPublish();
7:     end if
8: end while
9: while S_p ≠ NULL do
10:    TriggerPublish();
11: end while
```

معرفی روش

- تابع انتشار

Algorithm 2 *TriggerPublish*() .

```
1: Delete expiring K-anonymous clusters from S_k using ω;
2: Let t' be a tuple stored in buffer for δ(expiring tuple) and P' be a partition containing t';
3: if |P'| ≥ K then
4:     InPartitionClustering(t', P');
5: end if
6: if |S_p| ≥ K then
7:     MergeClustering(t', P');
8: else
9:     SingleAnonymization(t', P');
10: end if
```

- تابع خوشه‌بندی در پارتیشن

Algorithm 3 *InPartitionClustering*(t', P') .

```
1: Find K − 1 nearest tuples to t' from P' and form a virtual cluster C'_p;
2: Find K-anonymous cluster C_k from S_k defined by P' has minimum information loss;
3: if C_k ≠ NULL then
4:     if InfoLoss(C'_p) ≥ Infoloss(C_k) then
5:         Use cluster generalization of C_k to publish t';
6:         Remove t' from P'
7:         RETURN;
8:     end if
9: end if
10: Anonymize and publish all tuples of C'_p and remove published tuples from P';
11: Add C'_p to S_k;
```

معرفی روش

- فاصله بین دو چندتایی

$$Distance(t_1, t_2) = \frac{\sum_{q_i \in |Q_1 \cap Q_2|} d_i(q_i)}{|Q_1 \cap Q_2|}$$

$$d_i(q_i) = \begin{cases} \frac{|r_{i,1} - r_{i,2}|}{|R_{i,u} - R_{i,l}|} & \text{if } q_i \text{ is numerical} \\ \frac{|leaves(H_i)| - 1}{|leaves(DGH_i)| - 1} & \text{if } q_i \text{ is categorical} \end{cases}$$

- تابع ادغام خوشه‌ها

Algorithm 4 *MergeClustering*(t', P', R) .

Find $K-1$ cluster from S_k that can fully generalize t' with low information loss;
if $C_k \neq NULL$ **then**
    Use cluster generalization of C_k to publish t';
    Remove t' from P';
    RETURN;
end if
while $|P'| \geq K$ **do**
    Find non-empty partition P_{sim} from S_p which is most similar to P';
    Merge P_{sim} into P' and remove P_{sim} from S_p;
end while
Find $K-1$ nearest tuple to t' from P' and form a virtual cluster C'_m that has missingness;
Anonymize and publish C'_m;
Remove published tuples from P';
Re-assign remaining tuples of P' to respective partitions;

معرفی روش

- نحوه انتخاب خوشه جهت ادغام

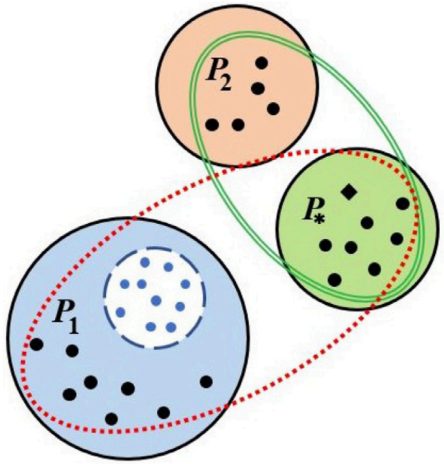
☑ مرحله اول: یافتن پارتیشن‌های شبیه با استفاده از فرمول:

$$Jaccard(P_1, P_2) = \frac{|Q_1 \cap Q_2|}{|Q_1 \cup Q_2|}$$

☑ مرحله دوم: انتخاب بهترین پارتیشن جهت ادغام با استفاده از فرمول شباهت:

$$Likeness(P, t, R) = \sum_{t_i \in P} Radar(t, t_i, R)$$

$$Radar(t, t_i, R) = \begin{cases} 1 & Distance(t, t_i) \leq R \\ 0 & Distance(t, t_i) > R \end{cases}$$



ارزیابی روش

- مشخصات ارزیابی

- با الگوریتم مشابه *FADS* و *IoTAnonymization* مقایسه شده است
- از مجموعه‌داده *Adult* و *PM2.5* برای ارزیابی عملکرد استفاده شده است
- مشخصات مجموعه‌داده‌های فوق به صورت زیر است:

QID descriptions of *Adult* dataset.

Attribute name	Type	Range	
		Min	Max
Age	*Numeric*	17	90
Final-weight	*Numeric*	13,769	1,484,705
Education-number	*Numeric*	1	16
Capital-gain	*Numeric*	0	99,999
Capital-loss	*Numeric*	0	4356
Hours-per-week	*Numeric*	1	99
		Hierarchy Height	**tree Nodes**
Education	*Categorical*	5	26
Marital-status	*Categorical*	4	11
Work-class	*Categorical*	5	13
Country	*Categorical*	4	62
Occupation	*Categorical*	3	15
Relationship	*Categorical*	3	7
Rage	*Categorical*	3	6
Gender	*Categorical*	2	3

QID descriptions of *PM2.5* dataset.

Attribute name	Type	Range	
		Min	Max
First-post	*Numeric*	1	1528
Second-post	*Numeric*	1	940
Third-post	*Numeric*	1	968
Dew-point	*Numeric*	−40	28
Temperature	*Numeric*	−25	41
Humidity	*Numeric*	2	100
Pressure	*Numeric*	975	1042
Wind-speed	*Numeric*	0	608
H-precipitation	*Numeric*	0	61.6
C-precipitation	*Numeric*	0	226.4
		Tree Height	**Nodes**
Season	*Categorical*	3	8
Wind-Direction	*Categorical*	2	5

ارزیابی روش

مشخصات ارزیابی

پارامترهای ارزیابی به صورت زیر است:

Algorithm name	Parameters
FADS	K=50, δ=2000,ω=200,α=0.001
IoT Anonymization	K=50, δ=2000, ω=200
K-VARP	K=50, δ=2000, ω=200, R=0.2

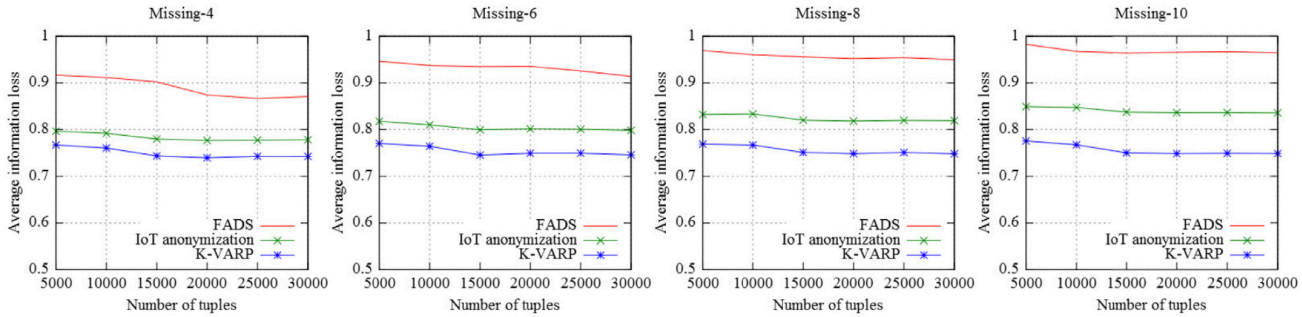
توصیف مجموعه‌داده (*Adult*):

Data size	Number of tuples with same number of missing values			
	Missing-0	Missing-1	Missing-2	Missing-3
5000	1250	1250	1250	1250
10,000	2500	2500	2500	2500
15,000	3750	3750	3750	3750
20,000	5000	5000	5000	5000
25,000	6250	6250	6250	6250
30,000	7500	7500	7500	7500

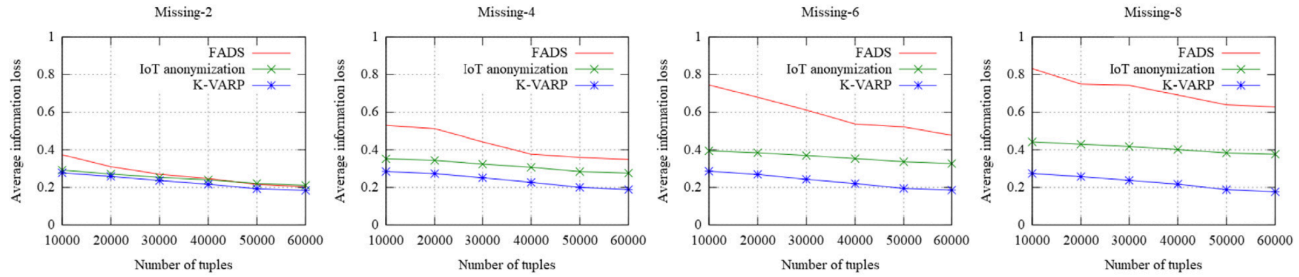
ارزیابی روش

- نتایج ارزیابی

☑ ارزیابی در مولفه‌های اتلاف اطلاعات، میزان استفاده مجدد خوشه‌ها، تعداد خوشه‌ها، میزان داده‌های حذف شده و زمان اجرا انجام شده است



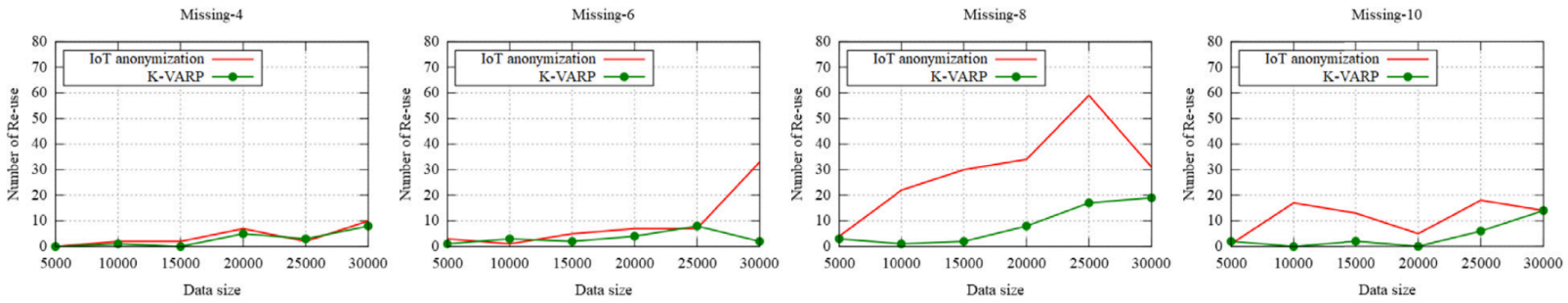
(a) Information loss (Adult dataset)



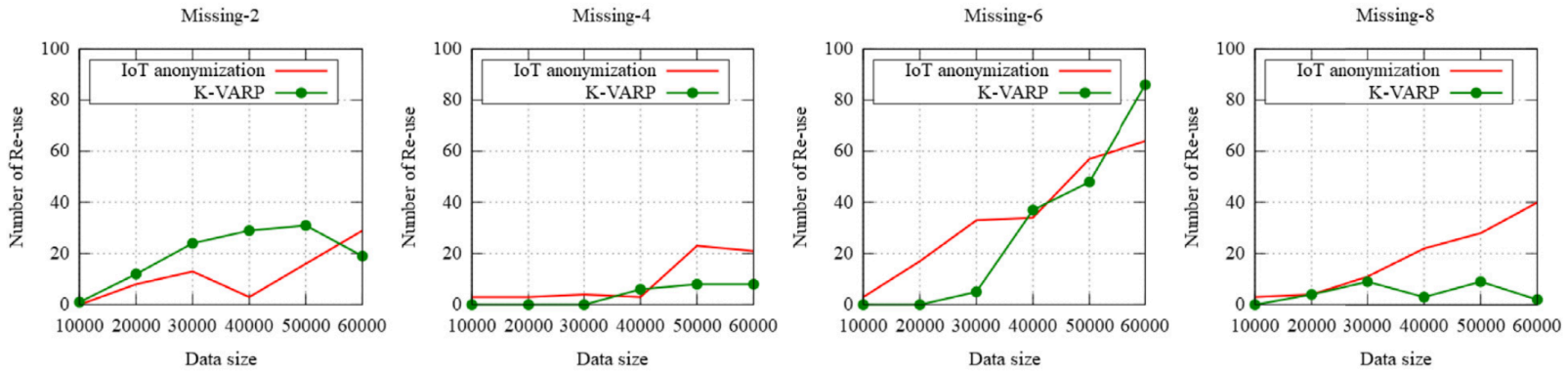
(b) Information loss (PM2.5 dataset)

ارزیابی روش

- نتایج ارزیابی (ادامه)



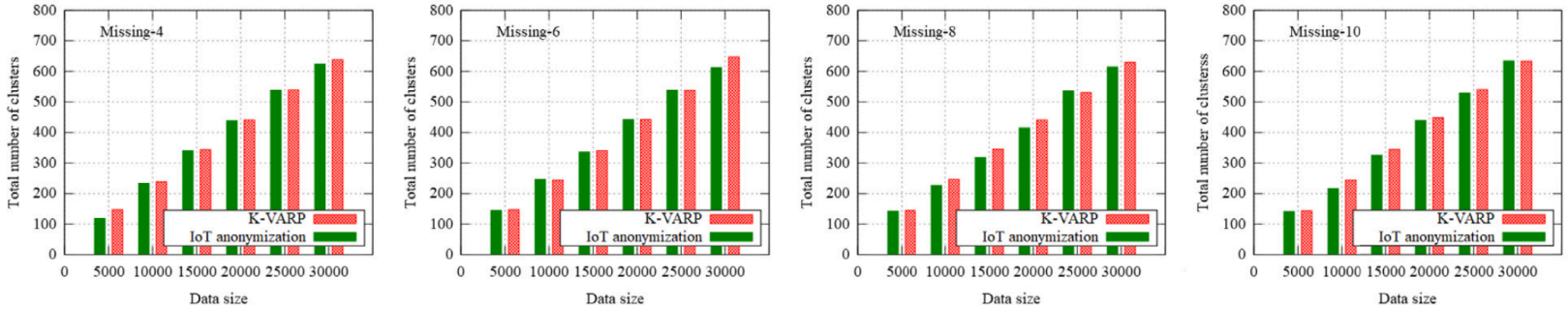
(a) Re-using (Adult dataset)



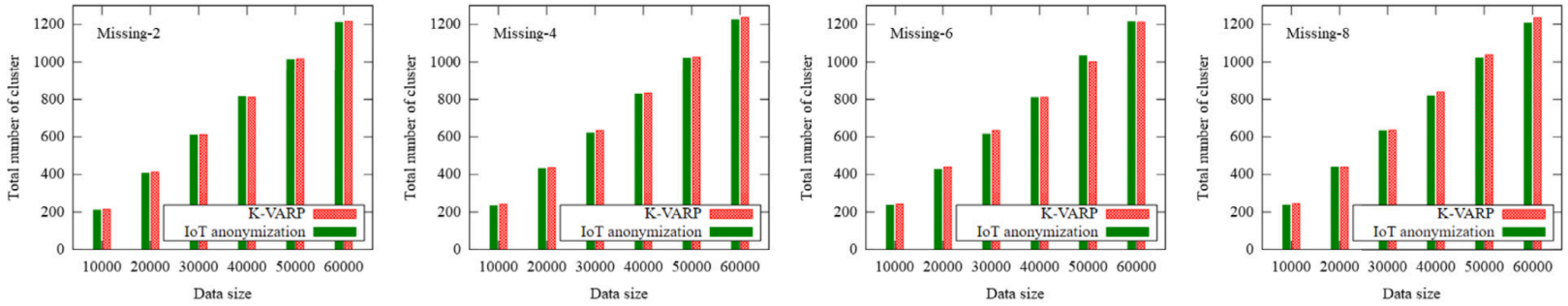
(b) Re-using (PM2.5 dataset)

ارزیابی روش

- نتایج ارزیابی (ادامه)



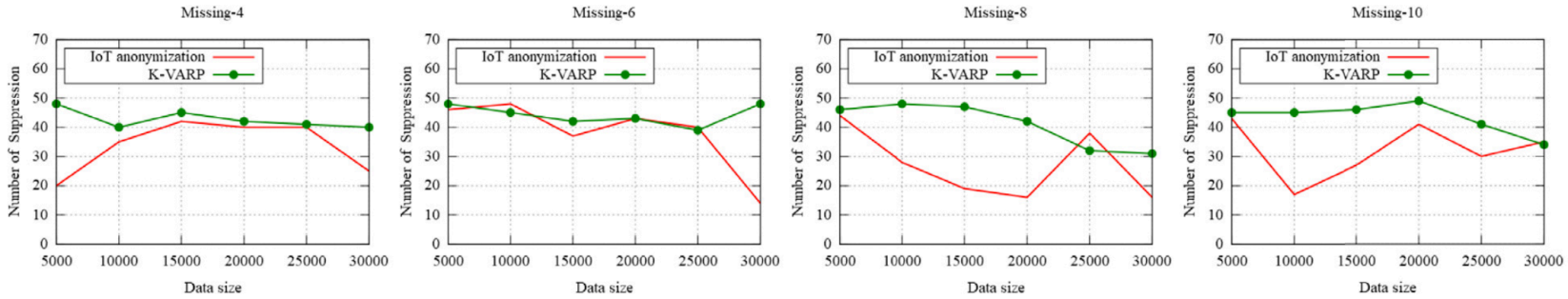
(a) Clusters created (Adult dataset)



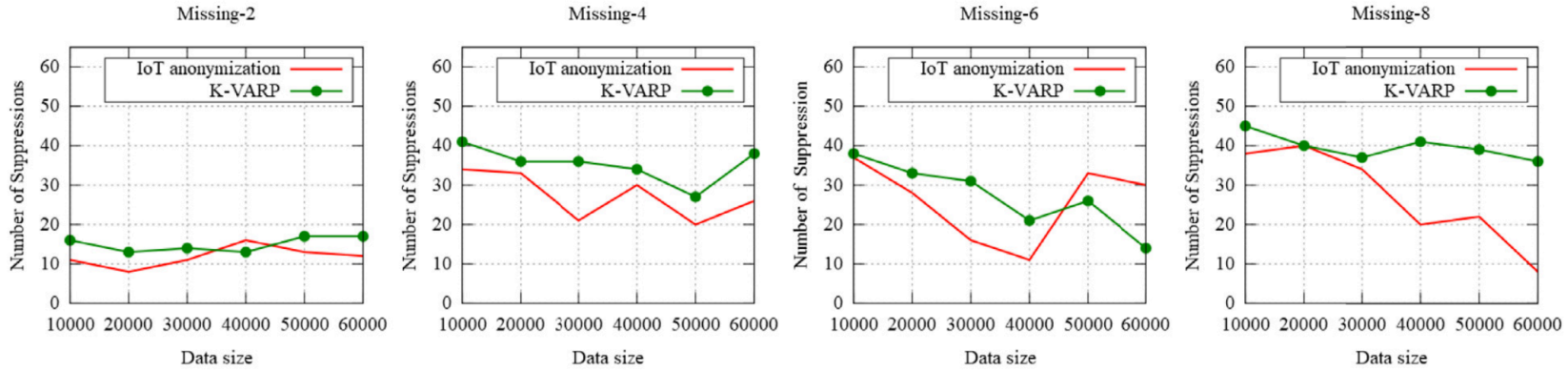
(b) Clusters created (PM2.5 dataset)

ارزیابی روش

- نتایج ارزیابی (ادامه)



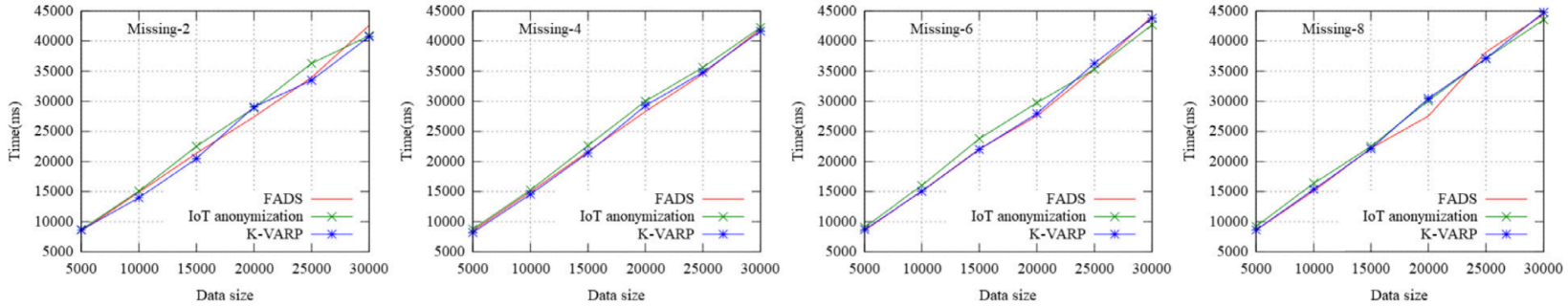
(a) Suppressions(Adult dataset)



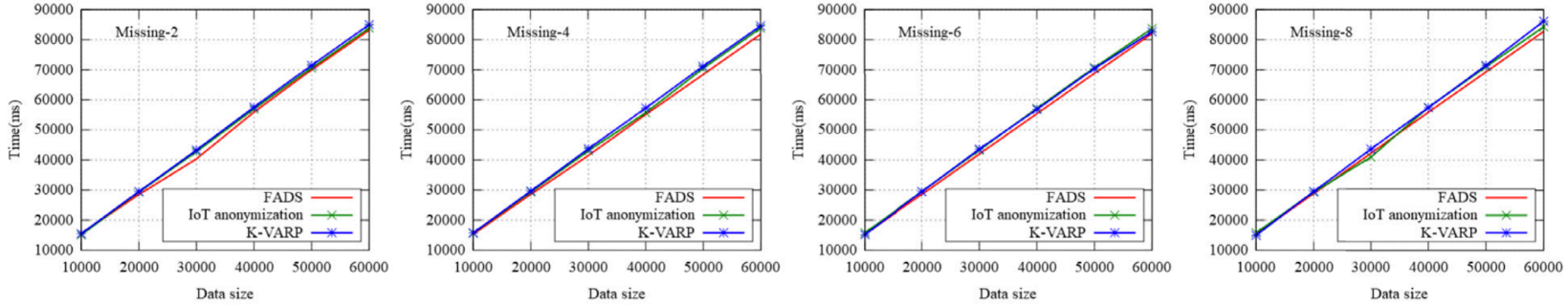
(b) Suppressions(PM2.5 dataset)

ارزیابی روش

- نتایج ارزیابی (ادامه)



(a) Runtime(Adult dataset)



(b) Runtime(PM2.5 dataset)

بررسی نقاط قوت و ضعف

- نقاط قوت
  - مناسب برای سیستم‌های واقعی مبتنی بر ابر (حاوی مقادیر گم شده)
  - استفاده از تابع شباهت برای پیدا کردن پارتیشن‌های مشابه جهت ادغام
- نقاط بهبود
  - بهینه‌سازی رویه ادغام خوشه‌ها

گزارش کار

- کارهای انجام شده
  - تولید داده‌های حاوی مقادیر گم‌شده از مجموعه‌داده اصلی به وسیله کد پایتون
  - توسعه و گسترش چارچوب فعلی جهت دریافت داده‌های حاوی مقادیر گم‌شده
  - پیاده‌سازی و تغییر مجموعه‌داده $Adult$ با توجه به تغییرات صفات شبه‌شناسه آن
- کارهای در حال انجام
  - پیاده‌سازی مفهوم استفاده مجدد از خوشه‌ها
  - گسترش چارچوب جهت دریافت داده‌های مجموعه‌داده $PM2.5$
  - اجرا و تهیه نمودار از نتایج ثبت شده

با تشکر از توجه شما