

عنوان ارائه:

یک روش خوشه‌بندی مبتنی بر تراکم برای حفظ حریم خصوصی  
در داده کاوی

**The Density-based clustering method for privacy preserving  
Data mining**

توسط: علیرضا صادقی نسب

استاد: دکتر حسین غفاریان

تاریخ ارائه: ۱۳۹۹/۹/۱۹

# مقدمه

## ■ اطلاعات مقاله

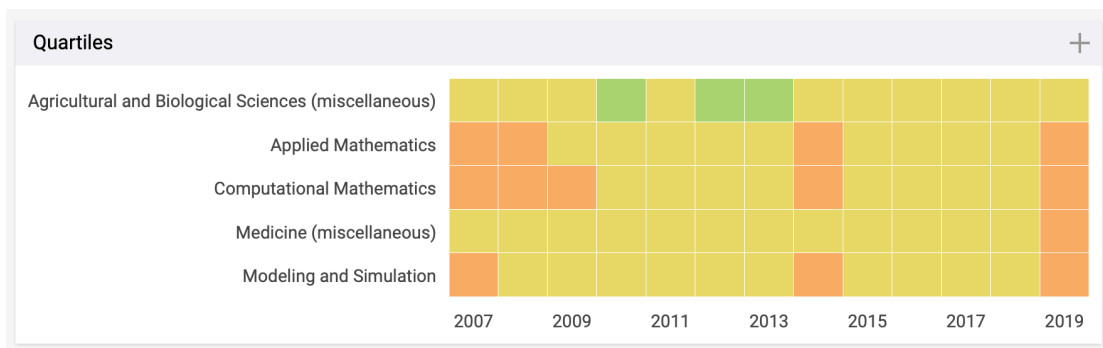
عنوان: *The density – based clustering method for privacy – preserving data mining*

سال چاپ: 2019

تعداد ارجاع: 8

مجله: *Mathematical Biosciences and Engineering*

ناشر: *American Institute of Mathematical Sciences*



## فهرست مطالب

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

# مقدمه

## ■ داده کاوی

- ★ فرآیندی برای کشف الگوها در مجموعه داده‌های بزرگ است
- ★ تقاطع حوزه‌هایی مانند یادگیری ماشین، آمار و سیستم‌های پایگاه داده است
- ★ یک حوزه بین رشته‌ای با هدف کلی استخراج اطلاعات غیربدیهی است
- ★ مرحله تجزیه و تحلیل فرآیند “کشف دانش در پایگاه داده” می‌باشد



# مقدمه

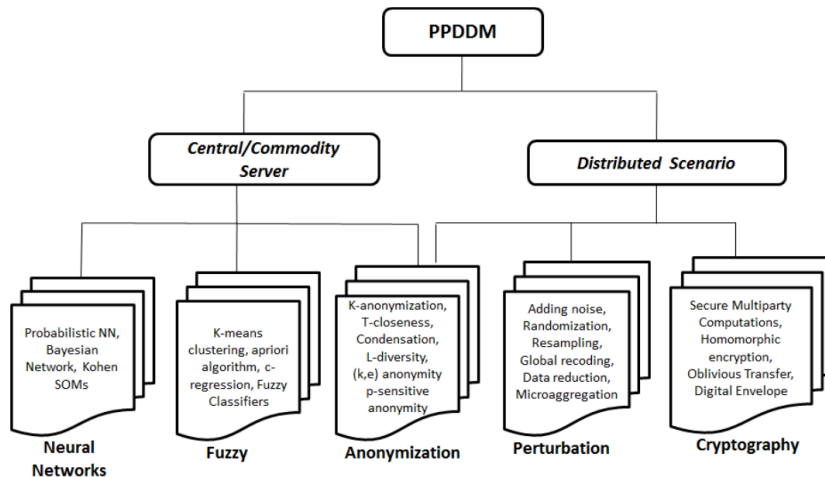
## ■ حفظ حریم خصوصی در داده کاوی

★ مجموعه اقداماتی است که در جهت رفع دغدغه‌های حریم خصوصی در فرآیند کاوش انجام می‌شود

★ یک راهکار حفظ حریم خصوصی دو ویژگی باید داشته باشد:

☑ اطلاعات و داده‌های شخصی و خصوصی را پنهان کند

☑ بتوان روی داده‌های دستکاری شده نیز، همان الگوهای مورد نیاز را به دست آورد



## مقدمه

■ فاکتورهای موثر در سودمندی حفظ حریم خصوصی

☑ شکست پنهان‌سازی: داده‌های حساسی که در انتها، پنهان نمی‌شوند

☑ هزینه مفقودی: داده‌های غیرحساسی که به اشتباه پنهان می‌شوند

☑ هزینه تصنعی: داده‌هایی که قبل از روش تکرارشونده نیستند ولی بعد از اعمال، تکرارشونده می‌شوند

★ فاکتورهای فوق می‌بایست توسط روش کمینه شده باشند، از این رو مسئله حفظ حریم خصوصی در داده‌کاوی یک مسئله *NP - hard* محسوب می‌شود



## مقدمه

▪ روش‌های محاسبات تکاملی جهت پیدا کردن بهینه‌سازی سراسری

🌐 الگوریتم ژنتیک: بهره‌گیری از نظریه داروین در مورد انتخاب طبیعی و بقا

✅ روش مبتنی بر جمعیت برای حل مسائل *NP – hard*

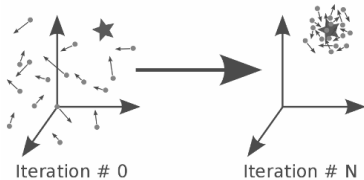
✅ از آپراتورهای جهش، کراس‌اُور و انتخاب برای پیدا کردن بهترین جواب، بهره می‌برد

🌐 بهینه‌سازی ازدحام ذرات: الهام گرفته از فعالیت‌های هجوم پرنده‌گان

✅ هر ذره به عنوان یک راه‌حل بالقوه مطرح است

✅ هر پرنده، سرعت مشخصی دارد که برای نشان دادن جهت نسبت به پاسخ‌های دیگر به کار می‌رود

✅ هر ذره در هر تکرار، خود را با بهترین مقدار خود و بهترین مقدار سراسری، بروز می‌کند



$$v_i(t + 1) = w \times v_i(t) + c_1 \times r_1 \times (pbest_i - x_i(t)) + c_2 \times r_2 \times (gbest - x_i(t))$$

$$x_i(t + 1) = x_i(t) + v_i(t + 1)$$

## مقدمه

### ■ مسائل بهینه‌سازی چند هدفه

★ با عناوین بهینه‌سازی برداری، بهینه‌سازی چند صفتی و بهینه‌سازی Pareto نیز شناخته می‌شود

★ مربوط به مشکلات بهینه‌سازی ریاضی است که شامل بیش از یک تابع هدف برای بهینه‌سازی همزمان است

★ در هر جایی که دو یا چند هدف متضاد در تصمیمات بهینه‌سازی حضور دارند، به کار برده می‌شود؛ مانند اقتصاد، لجستیک و غیره

★ مثال‌های کاربردی:

☆ بالا بردن عملکرد خودرو و به حداقل رساندن مصرف سوخت و انتشار آلاینده آن

☆ بالا بردن بازدهی دارائی و به حداقل رساندن ریسک‌های سرمایه‌گذاری



# معرفی روش

تعریف مقدمات و معرفی مسئله

$I = \{i_1, i_2, \dots, i_m\}$  → یک مجموعه متناهی از  $r$  موارد متمایز در پایگاه داده  $D$

$D = \{T_1, T_2, \dots, T_n\}$  → یک مجموعه متناهی از تراکنش‌ها  
هر  $T_q \in D$  زیرمجموعه‌ای از  $I$  است و یک شناسه یکتا  $q$  به نام  $TID$  دارد

$SI = \{s_1, s_2, \dots, s_k\}$  → یک مجموعه متناهی از موارد حساس که توسط کاربر مشخص می‌شود

⚠ هر  $s_i$  یی که در  $SI$  است، خود زیرمجموعه  $T_q$  محسوب می‌شود

★ مقدار  $\mu$  حداقل آستانه پشتیبانی در  $D$  است

★ اگر  $sup(i_j) \geq \mu \times |D|$ ، آنگاه مجموعه موارد  $i_j$  به عنوان یک مجموعه مکرر تعریف می‌شود

## معرفی روش

✓ تعریف ۱: به ازای هر  $s_i \in SI$ ، تعداد تراکنش برای حذف مجموعه مورد را  $N_{s_i}$  می‌نامیم و به

$$N_{s_i} = \frac{\text{sup}(s_i) - \mu \times |D|}{1 - \mu}$$

صورت روبه‌رو تعریف می‌کنیم:

✓ تعریف ۲: بیشینه تعداد تراکنش‌های حذف شده در بین تمامی مجموعه‌موارد  $SI$  را  $MDT$  می‌نامیم و

$$MDT = \max\{N_{s_1}, N_{s_2}, \dots, N_{s_k}\}$$

به صورت روبه‌رو تعریف می‌کنیم:

✓ تعریف ۳: متغیر  $\alpha$  را تعداد مجموعه‌موارد حساسی که پنهان نشده‌اند، می‌نامیم و آن را به صورت زیر

$$\alpha = |SI \cup L'|$$

تعریف می‌کنیم:

✓ تعریف ۴: متغیر  $\beta$  را تعداد مجموعه‌موارد غیرحساس و تکرارشونده‌ای که به اشتباه پنهان شده‌اند،

$$\beta = |L - SI - L'|$$

می‌نامیم و آن را به صورت روبه‌رو تعریف می‌کنیم:

✓ تعریف ۵: متغیر  $\gamma$  را تعداد مجموعه‌موارد غیرتکرارشونده‌ای که بعد از عملیات پنهان‌سازی، تکرارشونده

$$\gamma = |L' - L|$$

شده‌اند، می‌نامیم و آن را به صورت روبه‌رو تعریف می‌کنیم:

## معرفی روش

★ صورت مسئله: مسئله حفظ حریم خصوصی در داده‌کاوی با حذف تراکنش براساس بهینه‌سازی ازدحام ذرات چند هدفه را یک مسئله کمینه‌سازی سه عارضه جانبی آن (بعد از اعمال فرآیند پنهان‌سازی) می‌نامیم و آن را به صورت زیر تعریف می‌کنیم:

$$f = \min[f_1, f_2, f_3]$$

☆ در اینجا،  $f_1, f_2, f_3$  به ترتیب  $\alpha, \beta, \gamma$  می‌باشند

★ در ابتدای کار، جهت به دست آوردن تراکنش‌های بهتر جهت حذف در فرآیند *PPDM*، پایگاه‌داده تمامی تراکنش‌هایی که هر یک از اعضای *SI* را دارند را تصویر می‌کند و پایگاه‌داده جدید  $D^*$  نامیده می‌شود. بنابراین تمامی تراکنش‌هایی که داده حساس دارند برای فرآیند حذف تصویر شده و به عنوان ذره کاندیدا در فرآیند تکامل شناخته می‌شوند

★ سپس الگوریتم پیشنهادی (*CMPSO*) به صورت تکرارشونده اجرا شده و سه عارضه جانبی یاد شده را ارزیابی می‌کند

## معرفی روش

★ در الگوریتم *CMPSO*، هر ذره می‌تواند به عنوان یک راه‌حل ممکن با بردارهای *MDT* نشان داده شود و هر بردار شناسه تراکنش است که تراکنش بالقوه برای حذف را نشان می‌دهد.

☆ لازم به ذکر است که یک بردار در یک ذره، می‌تواند مقدار پوچ داشته باشد

☆ فرمول‌های پیشرفت و بروزرسانی الگوریتم به صورت زیر تعریف می‌شوند:

$$v_i(t + 1) = (pbest - x_i(t)) \cup (gbest - x_i(t))$$

$$x_i(t + 1) = rand(x_i, null) + v_i(t + 1)$$


\* راه‌حل غیرمسلط: راه‌حلی هستند که یک سازش مناسب (بدون تخریب هیچ یک) بین همه اهداف ایجاد می‌کند

\* راه حل *Pareto*: راه حل یک مسئله چند هدفه با مجموعه‌ای از نقاط *Pareto* داده می‌شود که هر یک به ترکیبی

منحصر به فرد از مقادیر تابع هدف به دست می‌یابد. ویژگی این راه‌حل‌ها این است که به صورت همزمان نمی‌توان

آن‌ها را در تمامی معیارها بهبود داد، بدون اینکه حداقل یکی از آن‌ها بدتر شود

# معرفی روش

شبه کد الگوریتم پیشنهادی 

---

**Algorithm 1:** Designed CMPSO Algorithm

---

**Input:**  $D^*$ , the projected database;  $L$ , the set of large itemsets for evolution;  $SI$ , the set of sensitive information to be hidden.

**Output:**  $D'$ , the sanitized database;  $Pset$ , the set of Pareto solutions.

```
1 initial  $N$  Particles with  $MDT$  size;
2 for each particle  $p$  in  $N$  do
3   evaluate  $f(p) := [f_1(p), f_2(p), f_3(p)]$ ;
4   obtain the non-dominated solutions  $Pset$ ;
5 while termination criteria is not achieved do
6   Gbest_update( $Pset$ );
7   update  $pbest$ ;
8   for each particle  $p'$  in  $N(t + 1)$  do
9     evaluate  $f(p') := [f_1(p'), f_2(p'), f_3(p')]$ ;
10    update the non-dominated solutions  $Pset$ ;
```


---

\* خط اول: تولید  $N$  ذره به اندازه  $MDT$  \* خط دوم و سوم: ارزیابی هر ذره با استفاده از تابع تناسب

\* خط چهارم: پیدا کردن پاسخ‌های غیرمسلط \* خط ششم و هفتم: بروزرسانی مقادیر  $pbest, gbest$

\* خط هشتم: ارزیابی ذرات بروز شده جهت بروزرسانی پاسخ‌های غیرمسلط

# معرفی روش

شبه کد الگوریتم پیشنهادی 

---

**Algorithm 2:** Gbest\_update(*Paretos*)

---

**Input:** *Paretos*, a set of Pareto solutions; *minpts*, the minimum number of solutions; *r*, the radius of a cluster.

**Output:** *gbest*, a global best particle for the updating progress.

```
1 set  $i := 1$ ;  
2 for each  $p$  in Paretos do  
3   if  $\text{sizeof}(p, r) \geq \text{minpts}$  then  
4      $c_i \leftarrow p$ ;  
5      $i++$ ;  
6 for each  $c_t$ ,  $t := 1$  to  $i$  do  
7    $\text{prob}(p \in c_t) := \frac{1}{i} \times \frac{1}{\text{sizeof}(c_t)}$ ;  
8  $\text{gbest} := \text{rand}(\text{prob}(p \in \text{Paretos}))$ ;
```

---

\* خط سوم: بررسی شرط خوشه‌بندی (تعداد پاسخ‌های *Pareto* در شعاع *r* بیشتر از *minpts* باشد)

\* خط چهارم: تخصیص‌دهی به خوشه \* خط ششم و هفتم: اختصاص یک احتمال به هر یک از خوشه‌ها

\* خط هشتم: مقداردهی پاسخ سراسری به صورت مقاداری تصادفی از احتمال‌های به دست آمده

# ارزیابی روش

## مشخصات ارزیابی

★ الگوریتم پیشنهادی با الگوریتم‌های یک هدفه *cpGA2DT* و *PSO2DT* از نظر اثربخشی و کارایی مقایسه شده است

★ الگوریتم‌ها به زبان جاوا پیاده‌سازی و در ماشین تحت سیستم عامل ویندوز با مشخصات *IntelCorei7 – 6700* و *8GBRAM* اجرا شده‌اند

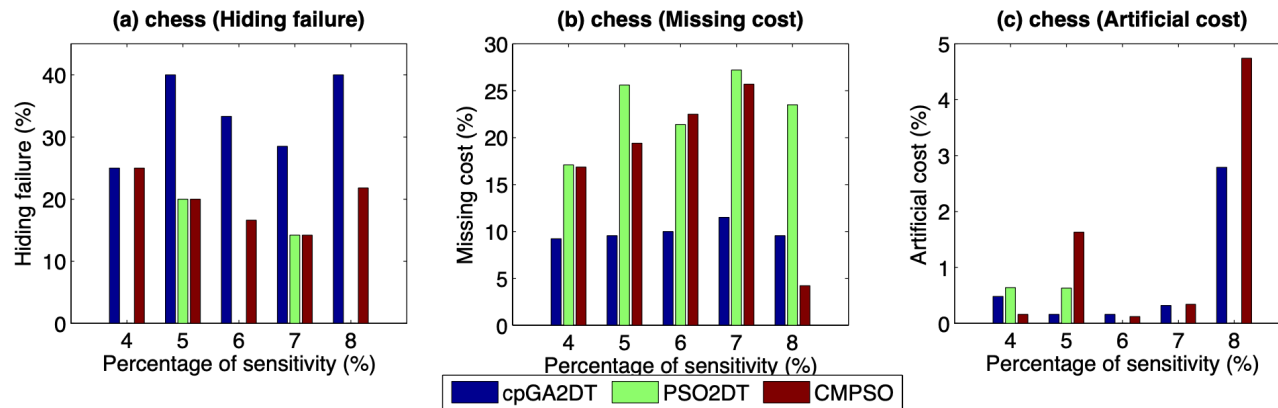
★ میزان جمعیت برای تمامی الگوریتم‌های تکاملی نامبرده شده، ۵۰ مقداردهی شده است

★ از آنجایی که یک الگوریتم چند هدفه، در انتها یک مجموعه *Pareto* تولید می‌کند؛ جهت مقایسه و ارزیابی عوارض جانبی، از میانگین پاسخ‌های تولید شده بهره برده شده است



# ارزیابی روش

☆ ارزیابی در مجموعه داده chess

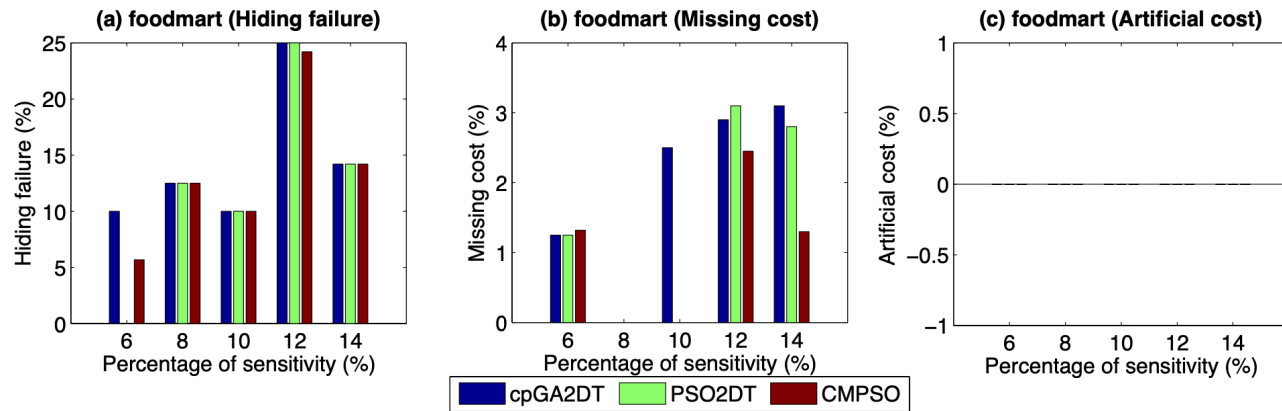


☆ علت اینکه در مجموعه داده‌ای مانند شطرنج، دو الگوریتم دیگر در بعضی حالات، نتیجه و عملکرد بهتری را نشان می‌دهند این است که همانطور که پیشتر گفته شد، الگوریتم پیشنهادی سه عارضه جانبی را از طریق میانگین پاسخ‌های به دست آمده ارزیابی می‌کند. در یک مجموعه داده چگال مانند شطرنج، تنوع راه‌حل‌های به دست آمده زیاد نیست؛ در نتیجه، پاسخ‌های *Pareto* به دست آمده، می‌توانند با هم همگرایی داشته باشند



# ارزیابی روش

☆ ارزیابی در مجموعه داده *foodmart*



☆ علت اینکه نتیجه در این حالت بهتر شده است این است که مجموعه داده فوق، پراکنده بود و در نتیجه، تنوع بیشتری از الگوریتم پیشنهادی به دست می آید. بنابراین، تراکنش های انتخاب شده، ممکن است حداقل عوارض جانبی را بعد از حذف داشته باشند

## بررسی نقاط قوت و ضعف

☆ نقاط قوت

✓ الگوریتم پیشنهادی مبتنی بر مسائل چند هدفه است

✓ عملکرد مناسب در برابر مجموعه داده‌های پراکنده

☆ نقاط ضعف

✓ عدم عملکرد مناسب در مواجهه با مجموعه داده‌های متراکم

## با تشکر از توجه شما