

MS rakendused: andmeanalüüs

Andmete klasterdamine

Klaster - ühesuguse käitumisega elementide rühm

Olgu antud 2 hulka X_1 ja X_2 .

Definitsioon 1. Hulgade X_1 ja X_2 lõikeks $X_1 \cap X_2$ nimetatakse elementide hulka $\{X_{ij}\}$, $X_{ij} \in X_1, X_2$.

Näide.

$$\begin{aligned} X_1 &= \{a, b, c\} & X_2 &= \{b\} \\ X_1 \cap X_2 &= \{b\} \end{aligned}$$

Olgu antud andmetabel $X(N, M) = \{X_{ij}\}$, $i=1, \dots, N$;
 $X_i = \{X_{ij}\}$, $j=1, \dots, M$; $X_{ij} = h_j = 0, 1, \dots, K_j - 1$.

Definitsioon 1. Lõikeks hulgal $X = \{X_i\}$ nimetatakse tunnuste j väärtuste hulka $\{X_{ij} = h_j\}$, $\{X_{ij} : \forall i, X_{ij} \in X_i\}$

Näide.

$X(2, 3)$.

$i \setminus j$	1	2	3
1	1	2	2
2	1	1	2

$\cap X = \cap_i X_i = \{1 * 2\}$, kus $*$ tähistab tühja elementi.

Kui $i=1$, siis $\cap X = X$,

Kui $i=0$, siis $\cap X = \emptyset$.

\Rightarrow Kui iga j korral $K_j=K$, siis $\max N=K^M$ ning lõikeid $(K+1)^M$

KLASSIKA

1. Algoritm

Leitakse lõiked üle objektipaaride, seejärel üle objektikolmikute, -nelikute,..., M-ikute.

- $2^N - 1$ võimalust lõikamiseks.

2. Algoritm

Leitakse lõiked üle objektipaaride. Seejärel lõiked üle lõigetepaaride. Seejärel lõiked üle saadud lõi-getepaaride...jne, kuni ühtegi uut lõiget ei lisandu.

- max iteratsioonide arv =
$$= \min \{M-1, \log_2(\max_j |h_j|)\}$$

3. Algoritm

```
FOR I1=1 TO N DO
  FOR I2=I1+1 TO N DO
    leia lõige
    IF lõige=∅ THEN NEXT I2
    FOR I3=I2+1 TO N DO
      leia lõige
      IF lõige=∅ THEN NEXT I3
      ...
      FOR IN=I(N-1)+1 TO N DO
        ...
      NEXT IN
    ...
  NEXT I2
NEXT I1
```

Prob leemid:

1) Kuidas vältida tühilõikeid?
2) ... korduvaid lõikeid?
3) Kuidas muuta lõigete leidmine juhitavaks?

MS lahendus

Idee.

Sagedus identifitseerib lõike elemendid üheselt.

Näide. $X(2,3)$.

$i \setminus j$	1	2	3
1	1	2	2
2	1	1	2
1	2	1	0
2	0	1	2

\Rightarrow Olgu antud tabel $X(N,M)$. Kui elemendi X_{ij} sagedus = N , kuulub element lõikesse.

MS: lõige on

1) tuum J. Mullati mõttes,

2) $\pi_{X \setminus \{X_{ij}\}}(X_{ij}) \leq \pi_X(X_{ij})$.

ALGORITM

S1. Leiame sagedused

S2. Kui sagedustabel tühi, väljasta lõige. $t=t-1$. Kui $t=-1$, mine LOPP. Valime juhtelemendi, lisame ta lõikesse, elimineerime ta

S3. Teeme väljavõtu, leiame sagedused

S4. Peegeldame eelmise taseme sagedusnullid

S5. Teeme tagasivõrdluse

S6. Kas elemendi sagedus = N_t ? Kui "jah", siis lisame elemendi lõikesse

S7. Mine S2

LOPP