Andmete maskeerimise meetodid: vahetus ja ümberpaigutus

Lauri Rätsep
Tartu Ülikool
a31218@ut.ee

Numbriliste andmete kaitse

- Numbrilised vs. Kategoorilised andmed
- Lihtne lahendus – andmete moonutamine

Probleem

- Kuidas muuta andmebaaside infot, nii et säiluksid selle statistilised omadused?
- Andmete vahetus
- Andmete ümberpaigutamine

Andmete vahetusalgoritm

- Klassipõhised vahetused
- Järjestatakse andmed mittekahanevalt
- Valitakse P(a)
- Vahetatakse väärtused a_i väärtusega a_i,nii et $|i-j|<P(a)*N/100$

C&S vahetusmeetod

- Andmehulk D = [S, X]
- $D = D^1 \cup D^2 = [S^1, X^1] \cup [S^2, X^2]$
- Järjestatakse X^1 ja X^2
- Suure n korral $X^1_j \sim X^2_j$
- Seetõttu vahetatakse omavahel X^1_j ja X^2_j
- Saadakse $[S^1, Y^1] \cup [S^2, Y^2] = D^{1*} \cup D^{2*} = D^*$

2DS ümberpaigutusalgoritm

Samm 1

- Genereeritakse originaalandmetest moonutatud valim $y=f_{X|S}\ (X|\ S = s_i)$
- Sama protsess iga kirje jaoks annab Y
- Praktikas on Y genereerimine heuristiline

2DS ümberpaigutusalgoritm [2]

Samm 2

- Järjestatakse nii originaalkirjed, kui ka uued genereeritud kirjed
- Genereeritud kirjed vahetatakse originaalkirjetega

Andmete kasutatavus

- Analüüside erinevused originaal ja maskeeritud andmete korral
- Mitu atribuuti
- Astakkorrelatsioon
- Valem:

$$\rho = 1 - \frac{6 \cdot \sum_{i=0}^{n} d_i}{n \cdot (n^2 - 1)}$$

Astakkorrelatsiooni näide

Id	Pikkused	Pikkuse järj.	Jalanumber	Jalanumbrite järj.	Vahed
1	1,67	1	37	1	0
2	1,90	5	45	4	1
3	1,81	3	43	3	0
4	1,73	2	39	2	0
5	1,89	4	46	5	1

- Astakkorrelatsioon on:

$$1 - \frac{6 \cdot 2}{5^3 - 5} = 0,9$$

Paljastusoht

- Isiksuse ja väärtuse paljastusoht
- Daleniuse, Duncani ja Lamberti definitsioon paljastusohule:
  - P({kirje seotakse isikuga| ligipääs olemas}) > P({kirje seotakse isikuga| ligipääs puudub})
  - Hinnangu viga SE({X&ligipääs olemas}) < SE({X&ligipääs puudub)

Eksperiment 1

- n = 30, 100, 300, 1000
- ρ = 0.00, 0.20 ,0.40 ,0.60 ,0.80 ,0.95
- Keskväärtus 0 ja dispersioon 1
- C&S meetod, ümberpaigutusmeetod, vahetusmeetod parameetritel P(a) = 10, 50, 100%
- Leiti 1000 korra astakkorrelatsiooni keskväärtus ja dispersioon.
- Arvutati korrelatsioon atribuutide X ja Y vahel.

Eksperiment 2

- Eesmärk hinnata isiku paljastusohtu
- k = 2, 3, 4, 5, 6 atribuuti, n = 30, 100, 300, 1000 kirjet
- samad algoritmid
- Fulleri meetod
- Eesmärk on vastandada olemasolevad väärtused maskeeritud väärtustega
- Leiti 1000 korra keskmine reidentifitseeritud kirjete arv.

Tulemused: andmete kasutatavus

Data set size	ρ		C&S (2002) method	Data shuffling	Data swapping (10%)	Data swapping (50%)	Data swapping (100%)
30	0.00	Bias	−0.000641	−0.001352	0.000551	−0.001045	−0.000950
		SE	0.066915	0.068458	0.108844	0.233072	0.263763
	0.20	Bias	−0.013962	−0.013758	−0.032443	−0.157984	−0.200020
		SE	0.066355	0.067250	0.107363	0.225624	0.259527
	0.40	Bias	−0.027262	−0.026777	−0.065604	−0.307165	−0.396881
		SE	0.063791	0.063063	0.101812	0.217107	0.245209
	0.60	Bias	−0.036869	−0.037706	−0.095700	−0.460840	−0.594337
		SE	0.057457	0.057452	0.093522	0.202012	0.224294
	0.80	Bias	−0.043728	−0.043730	−0.118208	−0.610875	−0.793003
		SE	0.046396	0.046919	0.079000	0.186656	0.201219
	0.95	Bias	−0.034909	−0.035300	−0.129913	−0.723215	−0.946191
		SE	0.030778	0.031542	0.062850	0.177532	0.189804
				...			
1,000	0.00	Bias	−0.000077	0.000031	0.000070	−0.001482	0.000286
		SE	0.002642	0.002622	0.015727	0.038503	0.042941
	0.20	Bias	−0.000823	−0.000745	−0.022539	−0.147750	−0.200895
		SE	0.002587	0.002598	0.014937	0.038802	0.042614
	0.40	Bias	−0.001332	−0.001518	−0.043130	−0.292616	−0.402256
		SE	0.000006	0.000006	0.000196	0.001266	0.001493
	0.60	Bias	−0.001884	−0.001976	−0.064120	−0.435326	−0.600270
		SE	0.002223	0.002264	0.012245	0.032524	0.035648
	0.80	Bias	−0.002268	−0.002264	−0.079229	−0.572165	−0.799787
		SE	0.001793	0.001676	0.010176	0.030401	0.032682
	0.95	Bias	−0.001740	−0.001675	−0.087542	−0.670493	−0.949409
		SE	0.000916	0.000884	0.007910	0.027450	0.031424

Tulemused: väärtuse paljastusoht

		Proportion of variability explained									
		C&S (2002) method		Data shuffling		Data swapping (10%)		Data swapping (50%)		Data swapping (100%)	
Data set size	ρ	$X_1 \mid Y_1$	$X_2 \mid Y_2$	$X_1 \mid Y_1$	$X_2 \mid Y_2$	$X_1 \mid Y_1$	$X_2 \mid Y_2$	$X_1 \mid Y_1$	$X_2 \mid Y_2$	$X_1 \mid Y_1$	$X_2 \mid Y_2$
30	0.00	0.934529	0.934854	0.000000	0.000001	0.828411	0.829045	0.215126	0.218040	0.001025	0.001145
	0.20	0.934529	0.934612	0.000000	0.000005	0.828411	0.829571	0.215138	0.217683	0.001025	0.001127
	0.40	0.934529	0.934200	0.000000	0.000000	0.828411	0.829046	0.215148	0.217049	0.001025	0.001121
	0.60	0.935262	0.935353	0.000003	0.000024	0.827240	0.828568	0.215452	0.216955	0.001467	0.001129
	0.80	0.935456	0.935104	0.000001	0.000001	0.829357	0.829117	0.215700	0.217574	0.001096	0.000835
	0.95	0.934534	0.934669	0.000003	0.000000	0.828431	0.828306	0.214481	0.217094	0.000955	0.001187
100	0.00	0.975144	0.974946	0.000000	0.000017	0.871473	0.871500	0.256386	0.257803	0.000108	0.000097
	0.20	0.975167	0.974961	0.000002	0.000001	0.871176	0.871048	0.256094	0.257500	0.000107	0.000121
	0.40	0.975144	0.974972	0.000001	0.000000	0.871473	0.871071	0.256386	0.257524	0.000108	0.000102
	0.60	0.975167	0.974993	0.000000	0.000005	0.871176	0.871274	0.256094	0.257529	0.000107	0.000126
	0.80	0.975139	0.974975	0.000000	0.000001	0.871086	0.871613	0.255945	0.257399	0.000113	0.000101
	0.95	0.975131	0.975073	0.000000	0.000000	0.871382	0.871811	0.256096	0.257585	0.000112	0.000100
300	0.00	0.990210	0.990387	0.000001	0.000000	0.882647	0.883192	0.266160	0.267111	0.000012	0.000006
	0.20	0.990230	0.990459	0.000004	0.000003	0.882726	0.883334	0.266230	0.267310	0.000012	0.000006
	0.40	0.990233	0.990371	0.000004	0.000001	0.882849	0.883464	0.266252	0.267737	0.000014	0.000006
	0.60	0.990213	0.990316	0.000006	0.000005	0.882670	0.883645	0.266144	0.267684	0.000013	0.000006
	0.80	0.990236	0.990177	0.000000	0.000002	0.882766	0.883278	0.266151	0.267329	0.000012	0.000006
	0.95	0.990222	0.990117	0.000000	0.000001	0.882731	0.882752	0.266178	0.267108	0.000013	0.000005
1,000	0.00	0.996733	0.996673	0.000000	0.000000	0.887913	0.886925	0.272723	0.271392	0.000000	0.000033
	0.20	0.996668	0.996644	0.000001	0.000009	0.887913	0.886911	0.272723	0.271262	0.000000	0.000032
	0.40	0.996650	0.996619	0.000000	0.000005	0.887913	0.887015	0.272723	0.271304	0.000000	0.000031
	0.60	0.996651	0.996584	0.000000	0.000000	0.887935	0.887288	0.272724	0.271375	0.000000	0.000032
	0.80	0.996690	0.996558	0.000000	0.000001	0.887943	0.887658	0.272744	0.271563	0.000000	0.000030
	0.95	0.996739	0.996637	0.000001	0.000000	0.887913	0.887864	0.272723	0.271674	0.000000	0.000030

Tulemused: Isiku paljastusoht

		Percentage of observations reidentified				
Data set size	Number of variables	C&S (2002) method (%)	Data shuffle (%)	Data swapping (10%)	Data swapping (50%)	Data swapping (100%)
30	2	72.42	3.55	60.04	26.61	4.29
	3	93.66	3.98	82.43	45.75	3.98
	4	98.57	4.20	91.05	56.06	4.03
	5	99.73	4.25	94.83	61.02	4.10
	6	99.97	3.89	96.66	61.82	4.10
100	2	66.46	1.03	55.32	10.37	1.14
	3	93.85	1.03	87.15	26.95	1.14
	4	99.01	1.06	96.56	50.32	1.21
	5	99.86	1.11	99.11	70.98	1.24
	6	99.96	1.01	99.61	84.31	1.21
300	2	62.62	0.32	48.68	3.90	0.38
	3	95.13	0.34	89.51	11.55	0.38
	4	99.44	0.35	98.33	27.26	0.47
	5	99.93	0.35	99.74	50.55	0.40
	6	99.99	0.32	99.94	72.40	0.42
1,000	2	61.18	0.11	37.01	1.33	0.11
	3	96.93	0.09	88.97	4.43	0.12
	4	99.66	0.10	98.63	11.72	0.12
	5	99.97	0.10	99.88	25.50	0.12
	6	100.00	0.12	99.99	46.07	0.13

Tulemused

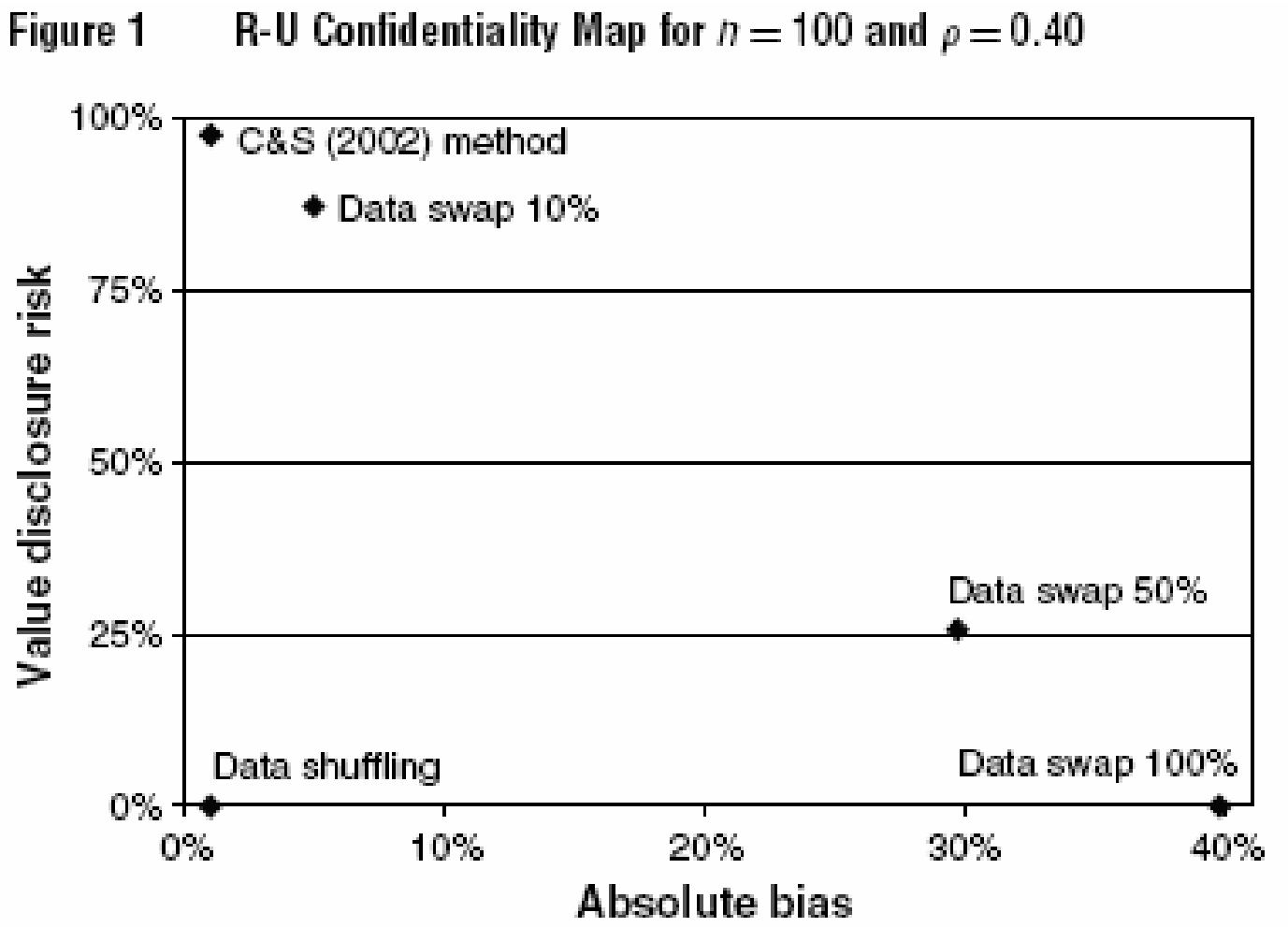
Figure 1 R-U Confidentiality Map for $n = 100$ and $\rho = 0.40$
100%
75%
50%
25%
0%
Value disclosure risk
C&S (2002) method
Data swap 10%
Data swap 50%
Data shuffling
Data swap 100%
0%
10%
20%
30%
40%
Absolute bias

Kokkuvõte

- C&S kasutatavuselt sama hea kui ümberpaigutusmeetod
- C&S meetodil suur isikupaljastusoht
- Vahetusmeetodi turvalisus hea, kui P(a) = 100%
- Ümberpaigutusmeetodi ilmselge paremus