Andmete maskeerimise meetodid: vahetus ja ümberpaigutus

Lauri Rätsep

1. mai 2007. a.

1 Sissejuhatus

Paljud organisatsioonid koguvad väärtuslikku infot klientide andmete analüüsimisest, kuid sellise konfidentsiaalse info levik võib rikkuda isiku privaatsust. Kategoorilised andmed on suhteliselt ohutud privaatsuse mõttes, kuid kuna numbrilised andmed võivad isiku väga üheselt määrata, siis viimaste aastatega on suurenenud just numbriliste andmete kaitse vajadus. Üldlevinud andmeturbe tehnikatest on selle probleemi lahendamiseks andmete moonutamine, mikroliitmine, andmete vahetus ja ümberpaigutamine.

Andmete vahetus ja ümberpaigutamine erinevad teistest lahendustest sellepoolest, et nad ei muuda originaalsete andmete väärtusi. Kuna ümberpaigutus on veel suhteliselt uus võrreldes andmete vahetusega, siis on tarvilik uurida nende kahe meetodi jõudlust ja turvalisust. Selles töös võrreldakse neid meetodeid tehisandmeid kasutades. Kõigepealt esitatakse nende tööpõhimõtted seejärel räägitakse turvalisusest ja lõpuks ka väljundandmete kvaliteedist.

2 Algoritmid

Andmete maskeerimise algoritme on palju, kuid enamik neist ei taga väljundandmete kvaliteeti - samad analüüsid originaalsetele ja maskeeritud andmetele erinevad liiga palju. Selle pärast peakski uurima algoritme, mis on nii turvalised kui ka konfidentsiaalsete tulemustega. Näiteks andmete moonutamine (*perturbation*) on üks meetod, mis muudab kirjete väärtusi nii, et säiliks nende kasutatavus. See algoritm on küll turvaline, kuid kahjuks võivad uued muudetud andmed kasutajatele kõlbmatuks osutuda. Selle tõttu tuuakse selles peatükis välja algoritmid, mis on turvalisuselt lähedased moonutamisega ning mille korral originaalandmete väärtuseid ei muudeta.

2.1 Reiss'i vahetusalgoritm

1978. aastal esitasid Tore Dalenius ja Steven Reiss vahetusalgoritmi salajaste atribuutide maskeerimiseks. See meetod on kõikide teiste vahetusalgoritmide

eelkäija, mis loodi eelkõige kategooriliste andmete maskeerimiseks. Kuna järgnev meetod on selle edasiarendus arvuliste andmete jaoks, siis lähemalt siinkohal Reiss'i algoritmi ei uurita [3].

2.2 Klassipõhine lähedaste vahetusalgoritm (*Rank-Based Proximity Swapping Algorithm*)

1996. aastal esitas Richard A. Moore siiani parima teadaoleva andmete vahetusalgoritmi (algoritm 1). Seda protseduuri saab rakendada vaid pidevatele suurustele. Erinevalt Reiss'i vahetusalgoritmist vahetatakse selle meetodiga omavahel vaid neid väärtusi, mis kuuluvad mingisse kindlasse klassi (vahetatavad suurused on arvuliselt lähedasemad). Sellised vahetused ei põhjusta andmetele väärtuste moonutusi [5].

Algoritm 1 Klassipõhine lähedaste vahetusalgoritm

1. Olgu andmete faili suurus N ja atribuutide arv M. Sorteeri andmed salajase atribuudi a järgi, nii et $i = 1, 2, ..., N$, kus $a_i \leq a_j$, $i < j$,
2. Määra väärtus $P(a)$, kus $0 \leq P(a) \leq 100$. Protseduuri idee on vahetada väärtus a_i väärtusega a_j, valides need kirjed nii, et protsentuaalne erinevus indeksite i ja j vahel ei oleks suurem kui $P(a)$ N-st. Ehk siis $|i - j| < P(a) * N/100$,
3. Väärtusta kõik ümardatud ja tühjade väärtustega kirjed lipuga "vahetatud". Kõik ülejäänud kirjed väärtusta lipuga "mittevahetatud".
4. Olgu madalaima "mittevahetatud" kirje indeks j. Vali juhuslik "mittevahetatud" kirje klassist intervalliga $[j+1, M]$, kus $M = min\{N, j + (P(a) * N/100)\}$. Oletame, et selle juhuslikult valitud kirje klassi indeks on k,
5. Vaheta väärtused a_j ja a_k. Määra nende kirjete lipuks "vahetatud",
6. Mine sammu 4 ja korda protseduuri, kuni kõik klassid on lipuga "vahetatud",
7. Oletame, et on vaja vahetada veel mingite väljade väärtusi (b, c, ...). Selleks pöördu tagasi sammu 1 ja korda protseduuri ühe välja kaupa uuesti. Seejuures võib kehtida, et $P(b) \neq P(a)$,
8. Kui vahetused on läbi, siis arvuta ja analüüsi statistilisi tulemusi. Kui need ei ole sobivas vahemikus, siis korda protseduuri uuesti, kasutades väiksemat väärtust $P(a), P(b), ...$

2.3 C&S meetod

2002. aastal esitasid Carlson ja Salabasis uue vahetusalgoritmi [4]. Oletame, et andmehulk $\mathbf{D}$ koosneb atribuutide hulkadest $\mathbf{S}$ (avalikud) ja $\mathbf{X}$ (salajased). Eeldame, et andmehulk $\mathbf{D}$ on juhuslikult jagatud kaheks (või rohkemaks) andmehulgaks $\mathbf{D}^1$ ja $\mathbf{D}^2$, mis sisaldavad vastavalt atribuutide hulkasid $[\mathbf{S}^1, \mathbf{X}^1]$ ja $[\mathbf{S}^2, \mathbf{X}^2]$. Kuna mõlemad andmehulgad on juhuslikud suurused, siis suure hulga väärtuste korral on $\mathbf{X}_j^1$ ligikaudselt võrdväärne hulgaga $\mathbf{X}_j^2$ (järjestades j-abil suurused). Seega hulk $\mathbf{Y}_j^1$ saadakse kui vahetatakse reastatud $\mathbf{X}_j^1$ väärtused $\mathbf{X}_j^2$ väärtustega ja hulk $\mathbf{Y}_j^2$ saadakse kui vahetatakse $\mathbf{X}_j^2$ väärtused $\mathbf{X}_j^1$ väärtuste-

ga. Seda protsessi korratakse iga j korral, et saada $\mathbf{Y}^1$ ja $\mathbf{Y}^2$. Seejärel seotakse maskeeritud atribuudid salajastega ($[\mathbf{S}^1, \mathbf{Y}^1]$ ja $[\mathbf{S}^2, \mathbf{Y}^2]$) ehk tekib kaks uut andmehulka $\mathbf{D}^{1*}$ ja $\mathbf{D}^{2*}$. Pärast seda on võimalik kombineerida need kaks andmehulka üheks $\mathbf{D}^*$.

2.4 2DS ümberpaigutusalgoritm

Erinevalt andmete vahetusest, kus kirjete i ja j salajased atribuudid vahetatakse omavahel, võib andmete ümberpaigutusalgoritmides küll i vahetuda j-ga, kuid j võib vahetuda k-ga jne.

Krish Muralidhar ja Rathindra Sarathy esitasid algoritmi 2DS (*The Two-Step-Shuffle*) [4], mis koosneb kahest osast. Esimeses osas genereeritakse moonutatud väärtused tinglikust jaotusest ning teises vahetatakse järjestatud esimese sammu tulemus originaalsete väärtustega.

Olgu $\mathbf{X}$ salajaste atribuutide hulk ning $\mathbf{S}$ avalike atribuutide hulk. Olgu$\boldsymbol{Y}^p$ *(perturbed)* juhuslik suurus moonutatud $\mathbf{X}$-st ja $\mathbf{Y}$ juhuslik suurus ümberpaigutatud $\mathbf{X}$-st. Esitagu $x_{i,j}$ ($x_{(i),j}$) i-ndat reastamata (reastatud) j-nda originaalatribuudi väärtust ning analoogiliselt $y^p_{i,j}$ ($y^p_{(i),j}$) i-ndat reastamata (reastatud) j-nda moonutatud atribuudi väärtust (algoritm 2).

Algoritm 2 2DS ümberpaigutusalgoritm

1. Genereeri maskeeritud väärtuste valim $\mathbf{y}^p_i$ tinglikust jaotusest $f_{X|S}(\mathbf{X}|\mathbf{S} = \mathbf{s}_i)$, kus $\mathbf{Y}^p$ja $\mathbf{X}$ on sõltumatud,
2. Korda protsessi iga vaatluse jaoks, et saada $\mathbf{Y}^p$,
3. Vaheta $y^p_{(i),j}$väärtusega $x_{(i),j}$, $(i) = 1, ..., N$, $j = 1, ..., M$ saavutamaks $\mathbf{Y}$,
4. Väljasta ümberpaigutatud andmed.

Praktikas on selle algoritmi realiseerimine raskendatud, kuna ühisjaotus $f(\mathbf{X}, \mathbf{S})$ ei ole alati teada. Kui ühisjaotus ongi teada, siis sellest $f_{X|Y}(\mathbf{X}|\mathbf{Y})$ tuletamine ei pruugi olla võimalik. Selle tõttu on algoritmi realiseerimine heuristiline, kasutades atribuutide ühis- ja tingliku jaotuse hinnanguid.

Ramesh A. Dandekar on samuti välja andnud ühe ümberpaigutusmeetodi LHS (Latin Hypercube Sampling). See meetod erineb eelmisest avalike atribuutide käsitlemise poolest erinevalt 2DS-ist, mis genereerib ümberpaigutatud salajaste atribuutide väärtused arvestades ka avalikke atribuute. Selle tõttu jäävad kehtima lisaks seostele salajaste atribuutide vahel ka seosed salajaste ja avalike atribuutide vahel. LHS-i peamine eesmärk on aga luua uus sünteetiline andmehulk ümberpaigutades väärtused terves andmebaasis, mille tõttu kahjustuvad seosed avalike ja salajaste atribuutide vahel. Seega 2DS on tunduvalt üldisem algoritm ning järgnev analüüs keskendub sellele ümberpaigutusmeetodile.

3 Algoritmide omadused

Selles peatükis selgitatakse meetmeid, mille abil võrreldakse algoritmide väljundandmete kasutatavust ning paljastusohtu.

3.1 Andmete kasutatavus

Andmete kasutatavus näitab, kui palju erinevad originaalsetele ja maskeeritud andmetele rakendatud analüüside tulemused. Kuna nii andmete vahetus kui ümberpaigutus jätavad ühe atribuudi korral andmete jaotuse samaks, siis oleks otstarbekam uurida nende meetodite omadusi mitme atribuudi korral. Seepärast on esimeseks andmete kasutatavuse mõõtmiseks astakkorrelatsioon. Kuna aga paljud andmete kasutajad tahaksid andmetele rakendada lihtsalt korrelatsiooni, siis oleks esmalt otstarbekam uurida kuidas erineb astakkorrelatsioon tavalisest korrelatsioonist [1].

Astakkorrelatsiooni arvutamiseks antakse esmalt mõlema atribuudi kõikidele väärtustele järjekorranumbrid ehk astakud. Seejärel arvutatakse astakute vahede ruudud d_i ning kasutatakse valemit 1.

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

Näiteks leiame astakkorrelatsiooni viie inimese pikkuste ja jalanumbrite vahel. Tabel 1 kirjeldab andmeid ning seal on leitud kasvavad järjestused pikkuse ja jalanumbri järgi ning nende järjekorranumbrite vahed. Kui nüüd kasutada valemit 1, siis saame

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

ehk astakkorrelatsioon inimeste pikkuste ja jalanumbrite vahel on tugev.

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

Tabel 1: Astakkorrelatsiooni näide

Et algoritmi poolt genereeritud andmed oleksid analüütiliselt kasutatavad, peab astakkorrelatsioon olema originaalsete ja maskeeritud andmete puhul sama tugev.

3.2 Paljastusoht

Paljastusoht jaguneb kaheks: isiku ja väärtuse paljastusohuks. Isiku paljastusoht on ründaja võime järeldada avalikest või maskeeritud väärtustest, et maskeeritud kirje kuulub teatud omanikule. Väärtuse paljastusoht on ründaja võime

järeldada kättesaadavatest väärtustest mingi salajase kirje väärtus. Selline deduktiivne paljastus võib aset leida ka siis, kui ligipääs maskeeritud andmetele puudub (avalike atribuutide järgi). Selle tõttu defineerisid Dalenius, Duncan ja Lambert paljastusohu, kui ligipääs maskeeritud andmetele on olemas. Nende definitsiooni järgi on paljastusoht suur:

1. kui tõenäosus, et identifitseerimata kirje seotakse isikuga ja ligipääs maskeeritud andmetele on olemas, on suurem kui tõenäosus, et identifitseerimata kirje seotakse isikuga ja ligipääs maskeeritud andmetele puudub.

2. kui salajase atribuudi hinnangu viga, juhul kui ligipääs maskeeritud andmetele on olemas, on väiksem, kui salajase atribuudi hinnangu viga, juhul kui ligipääs andmetele puudub.

Oletame, et ründaja proovib kõigepealt paljastada isikut ning seejärel mingit kindlat väärtust.

4 Eksperimendid

K. Muralidhar jt. [4] korraldasid 2 erinevat eksperimenti, et võrrelda vahetusalgoritme ja ümberpaigutusalgoritmi.

4.1 Eksperiment 1

Esimeses eksperimendis oli erinevaid kirjeid n $(= 30, 100, 300, 1000)$, mis genereeriti kahemuutujalisest normaaljaotusest kindla korrelatsiooni kordajaga ρ $(= 0.00, 0.20, 0.40, 0.60, 0.80, 0.95)$. Mõlema atribuudi korral olid keskväärtus ja dispersioon 0 ja 1. Seejärel maskeeriti andmed C&S meetodil, ümberpaigutusmeetodil ja siis vahetusmeetodil, kasutades kolme erinevat lähedusparameetrit $P(a)(= 10, 50, 100\%)$. Järgnevalt leiti astakkorrelatsioon nii originaalsete kui ka kõikide maskeeritud andmete puhul ning arvutati ja salvestati originaalsete ja maskeeritud andmete erinevused. Seda protsessi korrati 1000 korda ning arvutati keskmine erinevus ja dispersioon. Kogu protseduuri korrati iga n ja ρ puhul (tabel 2).

Sedasama eksperimenti kasutati ka väärtuse paljastusohu hindamiseks. Selleks arvutati ja salvestati korrelatsioon iga originaalatribuudi ning vastava maskeeritud atribuudi vahel. Atribuutide omavaheliste suhete paremaks kirjeldamiseks arvutati korrelatsiooni koefitsentide ruudud. Keskmine kirjeldus atribuutide suhetele arvutati kui 1000 tulemuse keskmine (tabel 4).

4.2 Eksperiment 2

Teise eksperimendi põhiline eesmärk oli kirjeldada isiku paljastusohtu. Selle tarbeks genereeriti mitmemuutujalisest normaaljaotusest n-elemendiline andmehulk, mis koosnes k atribuudist.

Sarnaselt eelmise eksperimendiga genereeriti viis hulka maskeeritud andmeid: üks ümberpaigutatusmeetodil ning neli vahetusalgortimidel. Kirjete omavahelised seoseid võrreldi Fuller'i meetodiga [2]. Oletame, et iga kirje korral on ründajal olemas mingid maskeerimata väärtused ning ta teab ka andmehulga üldiseid karakteristikuid. Ründaja eesmärk on vastandada olemasolevad väärtused maskeeritud väärtustega. Kõik kirjete maskeeritud väärtused, mis suudeti ära tunda, salvestati. Protsessi korrati 1000 korda ning leiti keskmine korduvalt identifitseeritud kirjete arv. Kõike seda korrati neljal väärtusel n $(= 30, 100, 300, 1000)$ ja viiel väärtusel k $(= 2, 3, 4, 5, 6)$ (tabel 5).

5 Tulemuste analüüs

5.1 Andmete kasutatavus

Tabelid 2 ja 4 kirjeldavad esimese eksperimendi tulemusi. Tabelis 2 on ära toodud iga n ja ρ korral astakkorrelatsiooni keskväärtus ja dispersioon originaalandmete ja kõikide maskeeritud andmete vahel. Nagu tabelist näha võib, on andmete ümberpaigutusmeetodil ning C&S algoritmil isegi väikese andmehulga $(n = 30)$ korral absoluutse keskväärtuse erinevus originaalsete ja maskeeritud andmete vahel väiksem kui 0.005. Sama suurus $n = 100$ korral on veelgi väiksem (väiksem kui 0.002) ning $n = 300, 1000$ korral peaaegu olematu. Ka dispersioon nendel juhtudel on väike (mitte kunagi üle 0.00912). Sellest võib järeldada, et kui kasutada C&S või ümberpaigutusalgoritmi väljundandmeid originaalsete asemel, siis on astakkorrelatsioon maskeeritud andmete puhul väga lähedane originaalsele.

Andmete vahetusmeetod nii heade omadustega pole. Isegi juhul kui vahetatavad kirjed on üksteisele väärtuselt väga lähedal $(P(a) = 10\%)$, võib märgata liiga suurt erinevust astakkorrelatsioonis originaalsete ja maskeeritud andmete vahel. Sest isegi kui $P(a) = 10\%$, $n = 30$ ja $\rho = 0.95$ on keskmine erinevus umbes -0.13. Loomulikult suurema $P(a)$ korral erinevad korrelatsioonid veelgi enam. Näiteks $P(a) = 100\%$, $n = 30$ ja $\rho = 0.95$ korral on keskmine erinevus -0.946. Nii suurte korrelatsioonide erinevuste põhjal võivad kasutajad arvata, et kahe atribuudi vaheline seos on väiksem kui originaalandmehulgas. Seega see eksperiment näitab, et kui hinnata andmete kasutatavust astakkorrelatsiooni abil, siis andmete ümberpaigutamine ja C&S on alati efektiivsemad kui vahetusmeetod mistahes parameetrite korral.

Tabelis 3 on ära toodud K. Muralidhar jt. [3] ühe varasema eksperimendi tulemused, kus sarnaselt tabeliga 2 on ära toodud eksperimendi tulemused, kasutades tavalist Pearsoni korrelatsioonikordajat. Carlson and Salabasis on näidanud, et andmehulga vahetamine algsega sarnase andmehulgaga (isegi kui see on sõltumatu), vähendab siiski korrelatsiooni. Seega arvatavasti ka mõlemad vaadeldavad meetodid suurendavad korrelatsioonide vahesid, kuigi andmebaasi kasvades peaks selline korrelatsioonide hääbumine vähenema. Selles tabelis kahjuks C&S vahetusmeetodit pole.

Ka selles tabelis on ümberpaigutusalgoritmil jällegi häid tulemusi ette näi-

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
100	0.00	Bias	−0.000107	−0.000260	−0.000989	−0.001064	−0.000050
		SE	0.022496	0.022872	0.050805	0.124219	0.142618
	0.20	Bias	−0.005065	−0.005090	−0.025164	−0.149044	−0.198753
		SE	0.022104	0.022358	0.050074	0.121077	0.140435
	0.40	Bias	−0.010030	−0.009880	−0.049637	−0.296934	−0.397648
		SE	0.021204	0.021248	0.047616	0.116631	0.133167
	0.60	Bias	−0.014155	−0.014178	−0.071887	−0.441791	−0.597483
		SE	0.019108	0.019010	0.042576	0.106818	0.120205
	0.80	Bias	−0.016616	−0.016775	−0.090536	−0.582097	−0.797881
		SE	0.015279	0.015314	0.036116	0.096673	0.106751
	0.95	Bias	−0.013024	−0.013062	−0.099643	−0.684039	−0.950754
		SE	0.008828	0.009120	0.028103	0.089905	0.102096
300	0.00	Bias	−0.000018	0.000340	0.000105	0.001819	0.001969
		SE	0.007890	0.007869	0.027860	0.069805	0.082571
	0.20	Bias	−0.002102	−0.001658	−0.022744	−0.146747	−0.199422
		SE	0.007752	0.007840	0.027705	0.067926	0.080740
	0.40	Bias	−0.004086	−0.003557	−0.045264	−0.291108	−0.398856
		SE	0.007396	0.007498	0.026525	0.066483	0.075039
	0.60	Bias	−0.005704	−0.005220	−0.065764	−0.434570	−0.599187
		SE	0.006826	0.006721	0.023619	0.059363	0.070232
	0.80	Bias	−0.006736	−0.006326	−0.082758	−0.575702	−0.800379
		SE	0.005494	0.005365	0.019968	0.054769	0.061630
	0.95	Bias	−0.005026	−0.004868	−0.091490	−0.674581	−0.948211
		SE	0.002974	0.002844	0.014765	0.051489	0.058183
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

Tabel 2: Esimese eksperimendi nihe ja standardviga algoritmide väljundandmete võrdlemiseks

data: kõikides andmehulkades suurustega 300 ja 1000 on korrelatsiooni hääbumine vaevumärgatav (-0.006). Suuruse 100 juures on suurim hääbumine siis, kui $\rho = 0.75$ (-0.017). Ilmselgelt on vähima andmete arvu 30 korral hääbumine maksimaalne (-0.043). Dispersioon on suhteliselt väike (< 0.005).

Vahetusmeetodite puhul saavutati paremad tulemused jällegi siis kui vahetusparameeter $P(a) = 10\%$. Korrelatsioonide erinevus on -0.130 ($n = 30, \rho = 0.95$) kuni -0.004 ($n = 1000, \rho = 0.05$). Kui $P(a) = 50\%$, siis on kõrgeim tulemus -0.722 ($n = 30, \rho = 0.95$) ja madalaim -0.034 ($n = 1000, \rho = 0.05$). Suurima parameetri $P(a) = 100\%$ korral on suurim -0.957 ($n = 30, \rho = 0.95$) ja vähim -0.050 ($n = 1000, \rho = 0.05$). Moore [5] on teinud nendest tulemustest palju järeldusi just vahetusmeetodi jaoks.

See tulemus näitab seega, et kasutajad valiksid meelsamini andmete ümberpaigutamise, kuna tabel 3 näitab, et andmete vahetusalgoritm tekitab liiga suure korrelatsioonide erinevuse originaalsete ja maskeeritud andmete vahel. Üleüldiselt võib öelda, et andmete ümberpaigutusalgoritmi ja C&S meetodi väljundandmed on vahetusmeetodi väljundist paremini kasutatavad.

5.2 Paljastusoht

Esimese eksperimendi tulemused väärtuse paljastusohu hindamiseks on kantud tabelisse 4 kahe atribuudi korrelatsioonikoefitsentide ruutudena (iga n ja ρ jaoks, et kirjeldada suhted originaalsete atribuutide ning maskeeritud atribuutide vahel). Kuna andmete ümberpaigutusalgoritm kasutab tinglikku jaotust, siis peaks see olema ka kõige turvalisem. Tabelist 4 selgubki, et maskeeritud atribuudi Y_i suhe originaalse atribuudiga X_i on väga nõrk. Isegi kui andmehulga suurus on väike ($n = 30$), jääb seoste suhet kirjeldav suurus alla 0.00003. Tabeli suuremate andmehulkade korral on see veelgi nõrgem.

Andmete vahetusmeetod aga nii head turvalisust ei paku. Kui $P(a) = 10\%$, siis suhe atribuutide vahel on väga kõrge. Kõikide andmehulkade suuruste juures on korrelatsioonikordaja ruut originaalsete ja maskeeritud andmete vahel alati suurem kui 0.82, kusjuures mida rohkem andmeid, seda suurem on kordaja. Seega ründaja võib väga täpse hinnangu saada kasutaja originaalandmete kohta. Ka $P(a) = 50\%$ korral suurenesid tulemused andmemahu kasvades, kuid kui $P(a) = 100\%$, siis enam mitte. Näiteks $n = 30$ on tulemused alati üle 0.001, kuid $n = 1000$ korral on kõik tulemused alla 0.00004.

C&S meetod on veelgi suurema väärtuse paljastusohuga. Kõik tulemused on üle 0.93 ning $n = 300$ ja $n = 1000$ korral isegi üle 0.99, mis tähendab, et maskeeritud atribuudid on originaalsetega väga tugevas seoses.

Seega võib öelda, et suhe atribuutide X_i ja Y_i vahel on alati nõrgim ümberpaigutatud andmete korral. Ümberpaigutusmeetod ei anna ründajale praktiliselt üldse infot originaalandmete kohta, mistõttu väärtuse paljastusoht on väiksem kui vahetusmeetodi korral. C&S meetod ei paku peaaegu mitte mingit turvalisust väärtuse paljastusohu vältimiseks.

Isiku paljastusohu määramiseks kasutati Fulleri meetodit [2], kus hinnati nende kirjete arvu, mille korral maskeeritud väärtus seostati originaalväärtusega. Vastavad tõenäosused on kantud tabelisse 5. Jällegi võib märgata ümber-

Dataset Size	ρ		Data Shuffling	Data Swapping (10%)	Data Swapping (50%)	Data Swapping (100%)
30	0.05	Average	-0.004	-0.016	-0.050	-0.054
		Variance	0.004	0.012	0.059	0.075
	0.25	Average	-0.017	-0.045	-0.191	-0.262
		Variance	0.004	0.012	0.052	0.065
	0.50	Average	-0.029	-0.084	-0.393	-0.493
		Variance	0.004	0.010	0.043	0.055
	0.75	Average	-0.044	-0.116	-0.564	-0.738
		Variance	0.003	0.007	0.034	0.046
	0.95	Average	-0.036	-0.130	-0.723	-0.957
		Variance	0.001	0.004	0.031	0.036
100	0.05	Average	-0.002	-0.008	-0.040	-0.050
		Variance	0.000	0.003	0.015	0.020
	0.25	Average	-0.006	-0.031	-0.180	-0.253
		Variance	0.000	0.003	0.016	0.019
	0.50	Average	-0.013	-0.063	-0.368	-0.497
		Variance	0.000	0.002	0.012	0.017
	0.75	Average	-0.017	-0.087	-0.543	-0.747
		Variance	0.000	0.002	0.009	0.012
	0.95	Average	-0.013	-0.100	-0.686	-0.948
		Variance	0.000	0.001	0.008	0.010
300	0.05	Average	-0.001	-0.006	-0.037	-0.052
		Variance	0.000	0.001	0.005	0.006
	0.25	Average	-0.003	-0.029	-0.182	-0.250
		Variance	0.000	0.001	0.005	0.007
	0.50	Average	-0.005	-0.056	-0.362	-0.500
		Variance	0.000	0.001	0.004	0.005
	0.75	Average	-0.007	-0.079	-0.538	-0.748
		Variance	0.000	0.000	0.003	0.004
	0.95	Average	-0.005	-0.090	-0.676	-0.955
		Variance	0.000	0.000	0.003	0.003
1000	0.05	Average	0.000	-0.006	-0.037	-0.050
		Variance	0.000	0.000	0.002	0.002
	0.25	Average	-0.001	-0.028	-0.181	-0.248
		Variance	0.000	0.000	0.001	0.002
	0.50	Average	-0.002	-0.054	-0.359	-0.499
		Variance	0.000	0.000	0.001	0.002
	0.75	Average	-0.002	-0.076	-0.537	-0.750
		Variance	0.000	0.000	0.001	0.001
	0.95	Average	-0.002	-0.088	-0.669	-0.951
		Variance	0.000	0.000	0.001	0.001

Tabel 3: Eksperimendiga 1 analoogilise eksperimendi tulemused kasutades Pearsoni korrelatsioonikordajat

		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

Tabel 4: Väärtuse paljastusoht

paigutusmeetodi paremust. Enamikel juhtudel on tõenäosus, et ümberpaigutatud andmetes leidub ära tuntud kirjeid, lähedane juhuslikule valimisele andmehulgast $(1/n)$. Näiteks kui $n = 30$, siis kõik tõenäosused jäävad 4% lähedale, samas juhusliku valimise tõenäosus on $3,33\%$. $n = 1000$ korral on kirjete äratundmise tõenäosus 0.09 kuni 0.12%, kusjuures juhusliku valimise tõenäosus on 0.10%. Seega võibki üldistada, et ümberpaigutusmeetodi korduva identifitseerimise tõenäosus ühtib juhusliku valiku tõenäosusega.

Andmevahetusmeetodil on märksa suurem isiku paljastusoht. Kui $P(a) = 10\%$ ja salajasi atribuute on 6, siis võib ründaja identifitseerida 96% ($n = 30$) kuni $99,99\%$ ($n = 1000$) kirjet. Isegi ainult kahe salajase atribuudiga on võimalik ründajal ära tunda 37 kuni 60% kirjetest. Juhul kui $P(a) = 50\%$ on tulemused märksa paremad ning $P(a) = 100\%$ korral on need peaaegu sama head kui ümberpaigutusalgoritmil.

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

Tabel 5: Äratuntud maskeeritud kirjete protsent

C&S meetodil on ülisuur isikupaljastusoht. Kui atribuute on vaid 2, saab siiski ära tunda üle 60% kirjetest. Kuue atribuudi korral võib peaaegu alati kõikide kirjete originaalväärtused viia kokku maskeeritud väärtustega.

Seega nii väärtuse kui ka isiku paljastusoht on igasuguse andmemahu ning atribuutide arvu juures väiksem kui kasutada ümberpaigutusmeetodit.

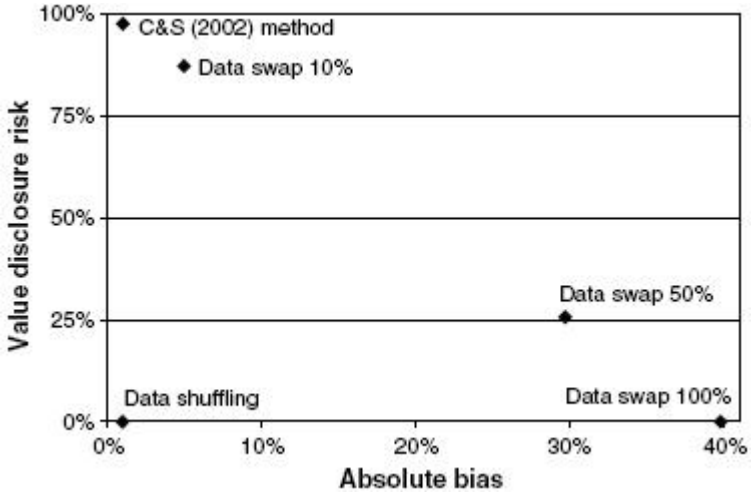
6 Tulemused

Isegi kui muuta parameetreid nagu kirjete arv, lähedusparameeter $P(a)$, salajaste atribuutide arv, jääb alati kehtima ümberpaigutusmeetodi üleolek vahetusmeetodi ja C&S meetodi ees. Joonisel 1 on ära toodud kõikide meetodite väärtuse paljastusohu ja andmete kasutatavuse võrdlus juhul kui $n = 100$, $\rho = 0.4$ [4].

Eksperimentide tulemustest võib veel järeldada, et ümberpaigutusalgoritmi korral:

1. kõik paarikaupa monotoonsed atribuutide vahelised seosed on originaalsete ja maskeeritud andmete puhul ühesugused,

2. ligipääsu lubamine maskeeritud andmetele ei suurenda paljastusohtu.



Joonis 1: Väärtuse paljastusohu (Value disclosure risk) ja nihke (bias) võrdlus

7 Kokkuvõte

Selle töö eesmärk oli uurida numbriliste andmete maskeerimist: andmete ümberpaigutust, C&S ja vahetusmeetodit. Töös toodi meetodite tööpõhimõtted ja nende võrdlemiseks kirjeldati kahte eksperimenti. Eksperimentidest järeldus, et andmed on paremini kasutatavamad ning vähem ohustatud nende paljastusest kui kasutada maskeerimiseks ümberpaigutusmeetodit. C&S väljundandmete kasutatavus on küll põhimõtteliselt sama hea kui ümberpaigutatud andmetel, kuid C&S meetod ei paku mitte mingit kaitset isiku paljastamisele. Seevastu vahetusmeetod (juhul kui $P(a) = 100\%$) pakub sama head turvalisust kui ümberpaigutusmeetod, kuid nii suure lähedusparameetri korral ei ole vahetusmeetodi väljundandmed enam kasutatavad.

Enne ümberpaigutusmeetodit sõltus andmete maskeerimiseks sobiva algoritmi valimine kasutaja vajadustest. Kui ta soovis, et salajased väärtused jääksid modifitseerimata, pidi ta kasutama eelpool kirjeldatud vahetusmeetodeid, mis jäid turvalisuselt ja väljundandmete kvaliteedilt alla andmete väärtusi moonutavatele meetoditele. Nüüd aga on kasutajatel võimalus andmeid maskeerida ümberpaigutusalgoritmiga nii, et kirjete väärtused jäävad samaks ning piisav turvalisus ja andmete kasutatavus on tagatud.

Viited

[1] Spearman's rank correlation coefitsent. `http://www.revision-notes.co.uk/revision/181.html` – viimati vaadatud 01.04.2007.

[2] W. A. Fuller. Masking procedures for microdata disclosure limitation. *Official Statist*, 9:383–406, 1993.

[3] Ramesh Dandekar Krish Muralidhar, RathindraŠarathy. Why swap when you can shuffle? a comparison of the proximity swap and data shuffle for numerical data. *Domingo-Ferrer and Franconi, Eds.*, LNCS 4302:164–176, 2006.

[4] RathindraŠarathy Krish Muralidhar. Data shuffling - a new masking approach for numerical data. *Managment Science*, 52:658–670, 2006.

[5] Richard A. Moore. Controlled data swapping for masking public use microdatasets. *Census Bureau Research Report 96/04*, 1996.