Heli Korkalainen

Päätöspuut yksityisyyden säilyttävässä tiedonlouhinnassa

Tietotekniikan kandidaatintutkielma

28. toukokuuta 2019

Jyväskylän yliopisto

Informaatioteknologian tiedekunta

Tekijä: Heli Korkalainen

Yhteystiedot: heli.k.korkalainen@student.jyu.fi

Ohjaaja: Sanna Mönkölä

Työn nimi: Päätöspuut yksityisyyden säilyttävässä tiedonlouhinnassa

Title in English: Decision trees in privacy preserving data mining

Työ: Kandidaatintutkielma

Sivumäärä: 20+0

Tiivistelmä: Yksityisyys voi vaarantua tiedonlouhinnan aikana, jonka seurauksena on syntynyt yksityisyyden säilyttävä tiedonlouhinta. Se tarkoittaa erilaisia tekniikoita ja menetelmiä, joilla voidaan louhia tietoa ilman, että arkaluonteisia tietoja paljastuu. Päätöspuita voidaan hyödyntää tiedonlouhinnassa tällaisissa menetelmissä. Tässä tutkielmassa tutkitaan sellaisia päätöspuumenetelmiä, jotka toimivat keskitetylle tai osapuolien kesken jaetulle datalle. Sopivan menetelmän valinta riippuu siitä, mitä ja miten tietoa louhitaan.

Avainsanat: tiedonlouhinta, yksityisyys, päätöspuu, yksityisyyden säilyttävä tiedonlouhinta

Abstract: Privacy can be at risk during data mining. Privacy preserving data mining has been created to answer this problem. Privacy preserving data mining is a set of techniques that can be used to mine data without revealing sensitive information. Decision trees can be used in these kinds of methods. The methods chosen for this study use either centralized or distributed data. Choosing the right method depends on the data that is used for mining.

Keywords: data mining, privacy, decision tree, privacy preserving data mining, PPDM

Kuviot

Kuvio 1. Esimerkki datan jaottelusta mukaillen Xu ym. (2014, s. 1160) 6

Sisältö

1 JOHDANTO 1

2 YKSITYISYYDEN SÄILYTTÄVÄ TIEDONLOUHINTA 3
2.1 Yksityisyyden säilyttävän tiedonlouhinnan vaiheita 4
2.2 Turvallinen yhteislaskenta 5

3 PÄÄTÖSPUIHIN PERUSTUVAT MENETELMÄT 7
3.1 Menetelmät keskitetylle datalle 7
3.2 Yhteislaskentaan perustuvat menetelmät 8

4 PÄÄTÖSPUIHIN PERUSTUVIEN MENETELMIEN VERTAILUA 10

5 YHTEENVETO 13

LÄHTEET 14

1 Johdanto

Viime aikoina yksityisyys on ollut pinnalla sekä GDPR-uudistuksen (Euroopan unioni) että erilaisten palvelujen, kuten Facebookin (New York Times 2018), tietovuotojen takia. Ihmiset ovat ymmärtäneet GDPR-uudistuksen myötä, kuinka paljon heistä kertyy tietoja lukemattomien eri palveluiden tietokantoihin. Lisäksi tietovuodot ovat herättäneet kysymyksen, kuinka paljon palvelut tietävät meistä ja kuinka helposti tiedot voivat päästä vääriin käsiin.

Ihmisistä kertyykin joka päivä paljon tietoa moniin eri tietokantoihin. Tätä dataa halutaan kerätä, koska sitä voidaan hyödyntää erilaisissa tarkoituksissa, kuten kaupallisissa hankkeissa tai tutkimuksissa. Esimerkiksi sosiaalisen median palvelut pyrkivät esittämään käyttäjilleen kohdennettuja mainoksia keräämiensä tietojen perusteella. Tällainen mainonta, ja muu samankaltainen dataan perustuva toiminta, tapahtuu tiedonlouhinnan avulla.

Tiedonlouhinta (engl. *data mining*) tarkoittaa tiedon tai mallien saamista suuresta määrästä dataa. Mallit voivat olla joko kuvaavia tai ennustavia, eli niillä voidaan kuvata tai ennustaa jotain ilmiötä. Koska tiedonlouhintaa voidaan soveltaa laajasti, voidaan näitä malleja käyttää monilla eri tavoilla. Tiedonlouhintaa hyödynnetään esimerkiksi lääketieteessä, päätöksenteossa ja mainonnassa.

Han, Kamber ja Pei (2011, s. 7–8) sekä Xu ym. (2014) jakavat tiedonlouhinnan eri vaiheisiin seuraavasti: Ensimmäisenä on datan esikäsittely, jossa datasta voidaan esimerkiksi poistaa kohinaa ja vääränlaista dataa. Lisäksi tässä vaiheessa valitaan sopiva data tiedonlouhintaa varten. Sen jälkeen data muutetaan sopivaan muotoon, jotta siitä voidaan louhia tietoa. Kolmantena on varsinainen tiedonlouhinta, jonka aikana käytetään erilaisia menetelmiä, joilla datasta saadaan selville hyödyllisiä tietoja. Lopuksi arvioidaan ja esitellään aiempien vaiheiden perusteella syntyneet mallit.

Tiedonlouhinnan aikana tai sen tuloksista voi kuitenkin paljastua arkaluonteisia tietoja, joilla tarkoitetaan sellaisia henkilökohtaisia tietoja, joita ei haluaisi muiden tietävän. Xu ym. (2014) määrittelevät arkaluontoisen tiedon tarkoittavan jotain sellaista tietoa, joka on saatavilla vain tietyille henkilöille ja jonka paljastuminen olisi vahingollista. Esimerkiksi jokin sairaus tai palkkataso voi olla tällaista tietoa. Yksityisyyttä ei kuitenkaan olisi tarpeellista loukata, sillä

tiedonlouhinnan tarkoituksena on asioiden yleistäminen, eikä yksilöiden tietojen paljastaminen (Malik, Ghazi ja Ali 2012).

Vastauksena tiedonlouhinnan yksityisyysongelmiin on syntynyt yksityisyyden säilyttävä tiedonlouhinta (engl. *privacy preserving data mining, PPDM*). Se tarkoittaa erilaisia tekniikoita, joilla datasta saadaan louhittua tietoa niin, ettei yksityisiä tai arkaluonteisia tietoja paljastu (Mendes ja Vilela 2017; Xu ym. 2014). Yksityisyyden säilyttävä tiedonlouhinta eroaa tiedonlouhinnasta pääosin vain yksityisyyden painottamisessa.

Yksityisyyden säilyttämiseen on kehitetty monia erilaisia menetelmiä, joita voidaan käyttää useassa tiedonlouhinnan vaiheessa. Koska tiedonlouhintaa voidaan soveltaa monella eri alalla, on metodeja paljon erilaisiin tiedonlouhinnan tehtäviin. Tässä tutkielmassa tarkastellaan sellaisia tiedonlouhintamenetelmiä, joiden toteutuksessa käytetään päätöspuita. Näitä menetelmiä käytetään tiedonlouhinnan kolmannessa vaiheessa eli varsinaisessa tiedonlouhinnassa.

Päätöspuulla tarkoitetaan yleisesti käytössä olevaa rakennetta, joka koostuu solmuista ja niiden lapsisolmuista. Jokainen puun solmu kuvastaa jotain ehtoa, jonka perusteella puurakenteessa liikutaan solmusta eteenpäin ehdon mukaiseen lapsisolmuun. Tiedonlouhinnassa päätöspuita voidaan käyttää luokittimina, sillä niiden rakenne sopii luokitteluun hyvin.

Tämä tutkielma on tehty kirjallisuuskatsauksena, sillä aiheesta on kertynyt laajasti kirjallisuutta Agrawalin ja Srikantin (2000) sekä Lindellin ja Pinkasin (2000) ensimmäisten töiden jälkeen. Päätöspuiden käyttäminen yksityisyyden säilyttävässä tiedonlouhinnassa on kiinnostanut monia tutkijoita, joten kirjallisuutta voi olla hyvä koota menetelmiin tutustumisen helpottamiseksi.

Tutkielma etenee seuraavasti: Luvussa 2 käsitellään yksityisyyden säilyttävää tiedonlouhintaa yleisesti. Sen jälkeen luvussa 3 käsitellään erilaisia päätöspuihin perustuvia yksityisyyden säilyttäviä menetelmiä. Näitä menetelmiä vertaillaan eri näkökulmista luvussa 4. Lopuksi luvussa 5 on yhteenveto käsitellyistä asioista.

2 Yksityisyyden säilyttävä tiedonlouhinta

Tässä luvussa tutustutaan yksityisyyden säilyttävään tiedonlouhintaan (engl. *privacy preserving data mining, PPDM*). Aluksi selvitetään yksityisyyden määritelmää, jonka jälkeen siirrytään yksityisyyden säilyttävään tiedonlouhintaan ja sen vaiheisiin. Lopuksi tutustutaan turvalliseen yhteislaskentaan, joka liittyy läheisesti moniin yksityisyyden säilyttäviin tiedonlouhinnan menetelmiin.

Yksityisyydelle on monia määritelmiä, ja arkikielessä sen merkitys eroaa usein tieteellisestä käytöstä. Clifton, Kantarcioglu ja Vaidya (2002) ovat määritelleet yksityisyyden säilyvän, jos henkilökohtaisia tietoja ei käytetä niin, että se vaikuttaa negatiivisesti jonkun elämään. Lisäksi he määrittelevät yksityisyyden tiedonlouhinnassa niin, että tiedonlouhinnassa ei mahdollisteta henkilökohtaisten tietojen väärinkäyttöä. Senosi ja Sibiya (2017) puolestaan määrittelevät yksityisyyden säilyttämisen olevan pahantahtoisen käyttäjän estämistä siten, ettei hän pysty tunnistamaan henkilön tietoja. Yksityisyyden säilyminen voidaan siis ajatella tietojen väärinkäytön estämisenä.

Yksityisyyden säilyttävä tiedonlouhinta voidaan ajatella tiedonlouhinnan laajennuksena, sillä se eroaa tiedonlouhinnasta pääasiassa vain yksityisyyden painottamisessa. Sen tarkoituksena on lisätä yksityisyyttä ja estää arkaluonteisten tietojen, kuten henkilötunnuksen, sairauden tai palkkatason, paljastumista. Clifton, Kantarcioglu ja Vaidya (2002) määrittelevät yksityisyyden säilyttävän tiedonlouhinnan olevan prosessi, jonka perusteella yksilöitä ei voida erottaa toisistaan. Tällaisesta tiedonlouhinnasta ei siis voi saada yksilön tietoja selville. Mendes ja Vilela (2017) määrittelevät yksityisyyden säilyttävän tiedonlouhinnan tarkoittavan kaikkia sellaisia menetelmiä, joilla voidaan louhia tietoa yksityisyyttä säilyttäen. Joistain tällaisista menetelmistä kerrotaan luvussa 3.

Erilaisia yksityisyyteen liittyviä menetelmiä onkin kehitetty paljon. Näiden menetelmien tarkoituksena on antaa jokin tietty määrä yksityisyyttä niin, että datan käytettävyys säilyy mahdollisimman hyvänä (Mendes ja Vilela 2017; Xu ym. 2014). Tämä on usein kuitenkin hankalaa, sillä yksityisyyttä lisättäessä tiedonlouhinnan tulokset heikentyvät ja menettävät tarkkuuttaan. Useimpien menetelmien tarkoituksena on joko suojata arkaluonteista dataa tie-

donlouhinnalta tai arkaluonteisia tiedonlouhinnan tuloksia (Malik, Ghazi ja Ali 2012). Toisin sanoen menetelmillä pyritään estämään arkaluonteisten tietojen paljastuminen tiedonlouhinnan tuloksista, mikä sopii yhteen Cliftonin, Kantarcioglun ja Vaidyan (2002) määritelmän kanssa. Arkaluonteista dataa ei myöskään pitäisi käyttää suoraan tiedonlouhinnassa (Xu ym. 2014), ettei yksilöllisiä tietoja voitaisi tunnistaa tiedonlouhinnasta.

2.1 Yksityisyyden säilyttävän tiedonlouhinnan vaiheita

Yksityisyyden säilyttävä tiedonlouhinta etenee samoin kuin perinteinen tiedonlouhinta. Vaiheet kuitenkin eroavat siten, että jokaisessa vaiheessa täytyy huomioida yksityisyys (Malik, Ghazi ja Ali 2012). Seuraavaksi kerrotaan, miten yksityisyyttä voidaan painottaa joidenkin tiedonlouhinnan vaiheiden aikana.

Datan kerääminen on tärkeä vaihe yksityisyyden säilyttävässä tiedonlouhinnassa, koska sillä voidaan vaikuttaa paljon yksityisyyden säilymiseen. Keräämisessä on tärkeää, että alkuperäisiä arvoja ei säilytetä, vaan niitä käytetään ainoastaan tallennettavien arvojen pohjana (Mendes ja Vilela 2017). Alkuperäisistä arvoista tulisi muokata tallennettavia arvoja jotenkin niin, että ne eivät enää vastaa oikeita arvoja. Jos alkuperäisiä arvoja ei säilytetä, datan päätyminen epäluotettavalle taholle ei ole niin suuri riski. Jokainen arvo pitäisi siis satunnaistaa yksitellen esimerkiksi lisäämällä kohinaa (Mendes ja Vilela 2017).

Yksityisyyden säilyttäminen on aiheellista myös silloin, kun dataa julkaistaan jollekin kolmannelle osapuolelle. Tätä yksityisyyden säilyttävän tiedonlouhinnan osa-aluetta kutsutaan myös yksityisyyden säilyttäväksi tiedon julkaisemiseksi (engl. *privacy preserving data publishing, PPDP*). Se tarkoittaa yksityisyyden säilyttämistä tiedon julkaisemisen aikana, jonka saavuttamiseksi käytetään pääasiassa datan anonymisaatiota (Mendes ja Vilela 2017; Xu ym. 2014).

Anonymisaatio perustuu identiteettien sekä arkaluonteisten tietojen piilottamiseen (Malik, Ghazi ja Ali 2012). Datasta poistetaan kaikki arvot, joiden pohjalta voidaan suoraan tunnistaa keneen ne liittyvät. Ongelmana ovat lisäksi epäsuorat tunnisteet (engl. *quasi-identifiers*), joista voidaan tunnistaa ihmisten henkilöllisyyksiä muiden taustatietojen avulla. Tällaisia arvoja tulisi muokata jotenkin ennen datan julkaisua (Xu ym. 2014). Anonymisaatioon käy-

tetään erilaisia tekniikoita. Arvoja voidaan yleistää, eli esimerkiksi tarkan iän voi muuttaa jollekin laajemmalle välille. Arvoja voidaan myös satunnaisesti poistaa, ja arat attribuutit sekä epäsuorat tunnisteet voidaan laittaa eri tauluihin (Mendes ja Vilela 2017). Dataa voidaan myös häiritä erilaisilla keinoilla, kuten kohinan lisäämisellä. Lueteltujen tapojen lisäksi on myös muitakin tapoja datan anonymisointiin.

Yksityisyyden säilyttävän tiedonlouhinnan loppuvaiheissa tärkeää on, ettei tiedonlouhinnan tuloksista paljastu arkaluonteisia tietoja. Tuloksista pitäisi siis poistaa arkaluonteiset tiedot. Esimerkiksi päätöspuita käyttävien menetelmien kohdalla luokittimen toimintaa voidaan heikentää, jos se paljastaa liikaa tietoja (Mendes ja Vilela 2017). Toiminnan heikentäminen kuitenkin heikentää myös ei-arkaluontoisten tietojen luokittelua, vaikka se auttaakin yksityisyyden säilyttämisessä.

2.2 Turvallinen yhteislaskenta

Turvallista yhteislaskentaa (engl. *secure multiparty computation, SMC*) käytetään joissain yksityisyyden säilyttävissä menetelmissä. Tähän laskentatapaan pohjautuvia menetelmiä on esitelty luvussa 3. Turvallisen yhteislaskennan tarkoituksena on laskea jotain yhdessä kahden tai useamman osapuolen kanssa. Laskentaan osallistujilla on yleensä kaikilla jonkin verran dataa, jota halutaan hyödyntää laskennassa muiden osallistujien datan kanssa. Osapuolet eivät kuitenkaan saa tietää muiden osapuolien dataa laskennan aikana, sillä turvallisen yhteislaskennan tarkoituksena on, että jokainen tietää vain oma datansa ja laskennan lopputuloksen (Xu ym. 2014). Turvallisessa yhteislaskennassa data on usein jaettu joko vaakasuoraan (engl. *horizontally partitioned data*) tai pystysuoraan (engl. *vertically partitioned data*) (Mendes ja Vilela 2017).

Datan jakamista on selvennetty kuviossa 1, jossa on esitetty datan jaottelua kuvitellulle tietokannan taululle. Vaakasuorasti jaotellussa datassa taulu on jaettu kahteen tauluun niin, että kumpaankin tauluun on tallennettu arvoja jokaisesta kerätystä asiasta (x1, x2, x3 ja x4). Arvot on kuitenkin jaettu taulujen kesken niin, että kummassakin on omia yksilöllisiä arvoja. Pystysuorasti jaotellussa datassa taulu on jaettu kahden taulun välille niin, että toiseen tauluun on tallennettu kaikki arvot joistain kerätyistä asioista (x1 ja x2). Vastaavasti toiseen

tauluun on tallennettu kaikki arvot joistain muista kerätyistä asioista (x3 ja x4). Usein taulun tiedot voi olla jaettu useammankin kuin kahden taulun välille.

Tietokanta

x1	x2	x3	x4
1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

Vaakasuorasti jaoteltu

x1	x2	x3	x4
1	2	3	4
5	6	7	8

x1	x2	x3	x4
9	10	11	12
13	14	15	16

Pystysuorasti jaoteltu

x1	x2
1	2
5	6
9	10
13	14

x3	x4
3	4
7	8
11	12
15	16

Kuvio 1. Esimerkki datan jaottelusta mukaillen Xu ym. (2014, s. 1160)

Turvallista yhteislaskentaa hyödynnetään yksityisyyden säilyttävässä tiedonlouhinnassa erilaisissa menetelmissä. Niissä oletetaan usein, että muut osapuolet ovat osittain rehellisiä (engl. *semi-honest adversaries*), eli he noudattavat vaadittua protokollaa, mutta voivat yrittää saada selville jotain ylimääräistä protokollan mukaisesti saaduista tiedoista (Mendes ja Vilela 2017). Osapuolet voivat myös olla pahantahtoisia (engl. *malicious adversaries*), eli he eivät noudata protokollaa ja yrittävät saada mahdollisimman paljon tietoa selville (Xu ym. 2014). Joitain esimerkkejä turvallisen yhteislaskennan käytöstä päätöspuuluokittimissa käydään seuraavaksi läpi luvussa 3.

3 Päätöspuihin perustuvat menetelmät

Tässä luvussa tutustutaan päätöspuihin perustuviin menetelmiin yksityisyyden säilyttävässä tiedonlouhinnassa. Menetelmät on jaettu tässä tutkielmassa kahteen luokkaan: sellaisiin, joissa louhitaan tietoa yhteen tietokantaan keskitetystä datasta ja sellaisiin, jotka lasketaan yhdessä kahden tai useamman osapuolen välillä turvallisen yhteislaskennan tapaan. Ensimmäisenä alaluvussa 3.1 käsitellään menetelmiä keskitetylle datalle, jonka jälkeen alaluvussa 3.2 käsitellään yhteislaskentaan perustuvia menetelmiä.

3.1 Menetelmät keskitetylle datalle

Agrawal ja Srikant (2000) kehittivät ensimmäisinä yksityisyyden säilyttävän päätöspuuluokittimen keskitetylle datalle. Menetelmässä päätöspuu luodaan syöttämällä ensin satunnaistamisfunktion läpi dataa. Sen jälkeen datan jakautuminen täytyy korjata vastaamaan alkuperäisen datan jakautumista, jotta arvot vastaisivat paremmin alkuperäistä dataa ja tiedonlouhinnan tulokset säilyisivät tarkkoina. Menetelmän satunnaistamisfunktio ei kuitenkaan toiminut, sillä alkuperäistä dataa pystyttiin palauttamaan menetelmän soveltamisen jälkeen (Kargupta ym. 2003). Menetelmää on kuitenkin käytetty muiden yksityisyyden säilyttävien päätöspuumenetelmien pohjana, kuten seuraavaksi käsiteltävässä Dowdin, Xun, ja Zhangin (2006) menetelmässä on tehty.

Dowd, Xu ja Zhang (2006) esittelevät päätöspuumenetelmän, joka perustuu satunnaisiin arvojen korvauksiin. Menetelmässä korvataan Agrawalin ja Srikantin (2000) satunnaistamisfunktio uudella datan satunnaistamistavalla, mutta muuten menetelmä pohjautuu heidän työhönsä. Aluksi Dowdin, Xun ja Zhangin (2006) menetelmässä dataa muutetaan korvaamalla arvoja satunnaisesti toisilla arvoilla. Korvaavat arvot valitaan alkuperäisen datan arvojen joukosta. Tämän jälkeen data muodostetaan uudestaan jakautumisen korjaamiseksi. Näin tiedonlouhinnassa ei käytetä suoraan alkuperäistä dataa, joten yksilökohtaisten tietojen tunnistaminen ei pitäisi olla mahdollista. Viimeisenä muokatusta datasta muodostetaan päätöspuu.

Fong ja Weber-Jahnke (2012) esittelevät menetelmän, jonka tarkoituksena on säilyttää yksityisyyttä muokkaamalla tiedonlouhinnassa käytettävää dataa ennen päätöspuun luomista.

Ensimmäisenä dataa muokataan niin, ettei se enää vastaa alkuperäistä dataa. Tämä muokattu data ei merkitse yksinään mitään, mutta kokonaisuutena sitä pystytään hyödyntämään. Tästä muokatusta datasta ei voi saada alkuperäisiä arvoja selville, ellei hyökkääjällä ole käytettävissä muokattu data kokonaisuudessaan. Muokkaamisen jälkeen datasta luodaan päätöspuu, jonka muodostaminen perustuu ID3-algoritmiin, josta kerrotaan lisää alaluvussa 3.2.

3.2 Yhteislaskentaan perustuvat menetelmät

Lindell ja Pinkas (2000) kehittivät ensimmäisinä yksityisyyden säilyttävän päätöspuumenetelmän kahden osapuolen välille. Menetelmässä tarkoituksena on, kuten muissakin yhteislaskennan menetelmissä, että toinen osapuoli ei saa tietää toisen tietokannasta ylimääräistä tietoa. Tässä oletuksena kuitenkin on, että osapuolet tietävät, mitä arvoja tietokantoihin on kerätty. Päätöspuuluokitin luodaan yhteisesti käyttämällä erilaisia yhteislaskennan keinoja. Menetelmä on muunnelma Quinlanin (1986) ID3-algoritmista, jota käytetään päätöspuuluokittimien luomiseen. ID3-algoritmissa puu luodaan rekursiivisesti ylhäältä alaspäin, ja lopputuloksena on suhteellisen yksinkertainen puu. Lindellin ja Pinkasin (2000) menetelmää on nykyään käytetty monien muiden menetelmien pohjalla.

Vaidya ja Clifton (2005) esittelevät menetelmän, jolla voidaan luoda päätöspuuluokitin pystysuoraan jaotellusta datasta kahden tai useamman osapuolen välillä. Se on vaikeampaa kuin vaakasuorasti jaotellun datan käsittely, sillä pystysuorasti jaotellussa datassa osapuolet eivät tiedä kaikkia laskennassa käytettäviä arvoja. Vaidyan ja Cliftonin (2005) menetelmä perustuu myös ID3-algoritmiin, jonka toimintaa on muokattu säilyttämään yksityisyyttä ja sopimaan yhteislaskentaan. Päätöspuuluokitinta muodostettaessa osapuolille paljastetaan vain puun rakenne ja jokaisessa solmussa olevan päätöksen tekijä. Päätöksen tekijän tietäminen on välttämätöntä, sillä luokittelussa edetään jokaisesta solmusta eteenpäin niin, että laskenta siirtyy sille osapuolelle, jonka päätöksessä ollaan.

Brickell ja Shmatikov (2009) esittelevät protokollan, joka toimii käyttäjän ja palvelimen välillä. Sen avulla käyttäjä voi luoda päätöspuuluokittimen palvelimella olevasta tietokannasta. Käyttäjä syöttää palvelimelle parametrit, joiden perusteella palvelin luo käyttäjälle päätöspuuluokittimen. Päätöspuu luodaan yksi kerros kerrallaan laskien yhtä aikaa koko kerroksen

solmujen tietoja. Palvelin ei opi mitään tehdystä luokittimesta, eikä käyttäjän antamista parametreista. Käyttäjä ei puolestaan saa tietää muuta kuin valmiin päätöspuuluokittimen, eikä mitään tietokannassa olevista arvoista. Menetelmä perustuu osittain Lindellin ja Pinkasin (2000) menetelmään yhteislaskennan tekniikoiden osalta. Brickellin ja Shmatikovin (2009) menetelmä eroaa muista yhteislaskennan menetelmistä siten, että tässä yhdellä osapuolella, eli palvelimella, on koko tietokanta. Menetelmää on hyödyllistä käyttää silloin, kun ei haluta paljastaa mitä tietokannasta louhitaan.

Vaidya ym. (2014) esittelevät satunnaisiin päätöspuihin perustuvan menetelmän, jolla voidaan rakentaa päätöspuita sekä vaakasuorasti että pystysuorasti jaotellusta datasta. Menetelmä perustuu Fanin ym. (2003) kehittämiin satunnaisiin päätöspuihin, jotka muodostetaan valitsemalla satunnaisesti arvoja puuhun. Vaidyan ym. (2014) menetelmässä on omat vaihtoehtonsa vaakasuorasti ja pystysuorasti jaotellulle datalle. Kun data on jaettu vaakasuorasti, jokainen osapuoli luo yksin rakenteen sovitulle määrälle satunnaisia päätöspuita, jonka jälkeen osapuolet valitsevat yhdessä sopivan rakenteen puulle. Sen jälkeen osapuolet laskevat yhdessä lehtisolmujen arvot. Kun data on jaettu pystysuorasti, osapuolet eivät voi yksin luoda päätöspuun rakennetta, sillä kaikki kerättävät arvot eivät ole yhden osapuolen tiedossa. Tämän takia puut pitää luoda yhdessä. Kummassakin tapauksessa puun rakenteen luomisen jälkeen osapuolet laskevat yhdessä lehtisolmujen arvot.

4 Päätöspuihin perustuvien menetelmien vertailua

Tässä luvussa tutkitaan luvussa 3 esiteltyjä menetelmiä eri näkökulmista. Monissa aiemmin esitellyissä menetelmissä on joitain eroja ja samankaltaisuuksia, joita on hyödyllistä huomioida. Ensimmäisenä kerrotaankin näistä eroista ja yhteneväisyyksistä, joiden jälkeen vertaillaan muutamien menetelmien toimintaa verrattuna sellaisiin menetelmiin, joissa yksityisyyttä ei ole säilytetty.

Alaluvussa 3.1 esitellyt menetelmät perustuvat datan häiritsemiseen (*data perturbation*). Agrawalin ja Srikantin (2000) sekä Dowdin, Xun ja Zhangin (2006) menetelmien ideana on häiritä dataa, jonka jälkeen sen jakauma korjataan vastaamaan alkuperäisen datan jakautumista. Tämän jälkeen muodostetaan päätöspuu. Fongin ja Weber-Jahnken (2012) menetelmä eroaa näistä siten, että siinä arvojen jakautumista ei tarvitse arvioida datan häiritsemisen jälkeen, sillä päätöspuun voi muodostaa suoraan menetelmän mukaisesti muokatusta datasta.

Yksityisyyden saavuttaminen dataa häiritsemällä on kuitenkin usein hankalaa, sillä joissain tilanteissa satunnaistamisen jälkeen datasta voi edelleen saada tietoja selville (Kargupta ym. 2003). Tämä on erityisesti ongelmana, jos datan arvot korreloivat keskenään jotenkin (Huang, Du ja Chen 2005). Esimerkiksi Agrawalin ja Srikantin (2000) menetelmästä pystyy saamaan tietoja selville datan satunnaistamisesta huolimatta. Dowd, Xu ja Zhang (2006) ovat kuitenkin huomioineet tämän ongelman, ja todistaneet, että heidän menetelmässään ei ole tätä ongelmaa.

Alaluvussa 3.2 olevat menetelmät perustuvat erilaisiin yhteislaskennan menetelmiin ja kryptografisiin tekniikoihin, joiden lisäksi menetelmissä on joitain erilaisia keinoja päätöspuiden luomiseen. Lindellin ja Pinkasin (2000) sekä Vaidyan ja Cliftonin (2005) menetelmät perustuvat kummatkin ID3-algoritmiin, jonka he ovat muuttaneet yksityisyyttä säilyttäväksi. Vaidyan ym. (2014) menetelmä puolestaan perustuu satunnaisiin päätöspuihin. Menetelmät eroavat toisistaan myös mahdollisten osapuolten lukumäärässä. Lindellin ja Pinkasin (2000) sekä Brickellin ja Shmatikovin (2009) menetelmissä osapuolten määrä on rajoitettu kahteen, kun taas Vaidyan ja Cliftonin (2005) sekä Vaidyan ym. (2014) menetelmissä osapuolia voi olla kaksi tai enemmän.

Alaluvun 3.2 menetelmissä samanlaista on se, että niiden tarkoituksena on paljastaa toisille osapuolille mahdollisimman vähän tietoa. Menetelmissä on kuitenkin tehty joitain kompromisseja tietojen paljastamisen suhteen, jotta laskenta ei ole liian vaativaa ja menetelmiä pystyttäisiin oikeasti käyttämään. Esimerkiksi Lindell ja Pinkas (2000) olettavat osapuolien tietävän mitä tietoja osapuolet ovat keränneet. Kaikissa alaluvussa 3.2 mainituissa menetelmissä myös oletetaan osapuolien olevan osittain rehellisiä, eikä pahantahtoisia osapuolia juuri käsitellä. Kuitenkin esimerkiksi Brickell ja Shmatikov (2009) mainitsevat, kuinka menetelmän pystyisi muuttamaan pahantahtoisia osapuolia varten, mutta se lisäisi laskentaa ja muuttaisi sitä raskaammaksi.

Menetelmistä olisi lisäksi hyödyllistä tietää, kuinka ne toimivat oikeissa tilanteissa. Joitain luvussa 3 esiteltyjä menetelmiä on testattu käytännössä ja selvitetty miten ne toimivat verrattuna sellaisiin menetelmiin, joissa ei ole säilytetty yksityisyyttä. Yksityisyyden säilyttäminen usein heikentää tarkkuutta ja lisää laskentaa, joten menetelmiä on hyödyllistä vertailla keskenään. Dowd, Xu ja Zhang (2006) ovat testanneet menetelmäänsä vertaamalla sillä luotuja puita suoraan alkuperäisestä datasta luotuun puuhun. Testistä saatiin selville, että menetelmällä luotujen puiden tarkkuus oli noin 75-80 % suoraan datasta luotuun puun tarkkuuteen verrattuna riippuen siitä, millä aineistolla menetelmää testattiin.

Fong ja Weber-Jahnke (2012) ovat myös testanneet menetelmäänsä kokeilemalla sen toimintaa erilaisilla datajoukoilla ja vertaamalla sitä sellaiseen puuhun, joka on luotu suoraan ilman yksityisyyttä säilyttäviä menetelmiä. Testauksesta he huomasivat, että menetelmällä luodut päätöspuut muokatusta datasta olivat yhtä tarkkoja kuin alkuperäisestä datasta luodut päätöspuut. Menetelmän kääntöpuolena kuitenkin on, että muokatun datan käyttäminen vie enemmän tilaa kuin suoraan alkuperäisen datan käyttäminen. Tilantarve on jopa noin kahdeksankertainen normaalin menetelmän tilankäyttöön verrattuna (Fong ja Weber-Jahnke 2012).

Vaidya ym. (2014) ovat tutkineet menetelmänsä toimintaa vertaamalla sitä Vaidyan ja Cliftonin (2005) menetelmään. Testauksesta huomattiin, että Vaidyan ym. (2014) satunnaisiin päätöspuihin perustuva menetelmä on nopeampi puiden luomisessa kuin Vaidyan ja Cliftonin (2005) yksityisyyttä säilyttävä ID3-algoritmin muunnelma. Vaidyan ja Cliftonin (2005) menetelmä vie kuitenkin noin kaksi kertaa enemmän aikaa päätöspuun luomisessa kuin al-

kuperäinen ID3-algoritmi (Vaidya ym. 2014), joten se on lähtökohtaisestikin hieman hidas. Satunnaiset päätöspuut ovat kuitenkin hitaampia luokittelussa verrattuna Vaidyan ja Cliftonin (2005) menetelmään, mikä voi johtua niiden satunnaisesta rakenteesta. Päätöspuiden tarkkuutta tarkasteltaessa huomattiin, että satunnaisten päätöspuiden tarkkuus on hieman heikompi kuin yksityisyyttä säilyttävillä ID3-puilla (Vaidya ym. 2014). Satunnaiset päätöspuut eroavat yksityisyyttä säilyttävistä ID3-puista silti nopeudellaan, joten ne voivat sopia paremmin käytännön sovelluksiin.

Menetelmien vertailu keskenään on kuitenkin hieman hankalaa, sillä ne ovat tarkoitettu käytettäväksi erilaisiin tilanteisiin. Hyvä esimerkki tästä on Brickellin ja Shmatikovin (2009) menetelmä, joka sopii käytettäväksi vain hyvin rajattuun tilanteeseen palvelimen ja käyttäjän välille. Yksityisyyden säilyttävän tiedonlouhinnan menetelmät ovatkin usein hyvin tapauskohtaisia, joten menetelmän valinta riippuukin täysin tilanteesta.

5 Yhteenveto

Yksityisyys on herättänyt viime aikoina huomiota tiedonlouhinnassa, koska ihmisistä kertyy yhä enemmän tietoja erilaisiin tietokantoihin. Erityisesti arkaluonteisten tietojen käyttäminen tiedonlouhinnassa on herättänyt huolta. Tiedonlouhinnassa ei kuitenkaan ole välttämätöntä loukata yksityisyyttä, koska sen tarkoituksena on yleistää ja saada selville yleisluontoista tietoa, eikä tarkoituksena ole tutkia yksilöiden tietoja. Yksityisyyden säilyttävää tiedonlouhintaa on kehitetty vastaamaan tähän ongelmaan. Yksityisyyden säilyttävä tiedonlouhinta tarkoittaa erilaisia tekniikoita, joilla datasta saadaan louhittua tietoa niin, ettei arkaluonteisia tietoja paljastu prosessin seurauksena.

Yksityisyyttä on mahdollista säilyttää käyttämällä erilaisia menetelmiä tiedonlouhinnan eri vaiheissa. Päätöspuita hyödynnettäessä käytetään usein menetelmiä keskitetylle datalle tai turvallisen yhteislaskennan menetelmiä useamman osapuolen kesken jaetulle datalle. Keskitetylle datalle luodut menetelmät usein häiritsevät dataa jollain tavalla, jonka jälkeen datan jakauma korjataan. Yhteislaskentaan luodut menetelmät puolestaan käyttävät erilaisia turvallisen yhteislaskennan menetelmiä ja pyrkivät paljastamaan mahdollisimman vähän tietoa muille. Menetelmiä on hyvin monia erilaisia, koska usein menetelmät sopivat vain johonkin tiettyyn tilanteeseen. Lisäksi valinnassa voidaan painottaa esimerkiksi menetelmän nopeutta, tiedonlouhinnan tulosten tarkkuutta tai yksityisyyden tasoa.

Lähteet

Agrawal, Rakesh, ja Ramakrishnan Srikant. 2000. “Privacy-preserving Data Mining”. *SIGMOD Rec.* 29 (2): 439–450. doi:10.1145/335191.335438. http://doi.acm.org/10.1145/335191.335438.

Brickell, Justin, ja Vitaly Shmatikov. 2009. “Privacy-Preserving Classifier Learning”. Teoksessa *Financial Cryptography and Data Security,* toimittanut Roger Dingledine ja Philippe Golle, 128–147. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 9783-642035494.

Clifton, Chris, Murat Kantarcioglu ja Jaideep Vaidya. 2002. “Defining privacy for data mining”. Teoksessa *National science foundation workshop on next generation data mining,* nide 1. 26.

Dowd, Jim, Shouhuai Xu ja Weining Zhang. 2006. “Privacy-Preserving Decision Tree Mining Based on Random Substitutions”, 3995:145–159. Emerging Trends in Information and Communication Security. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/11766155_11.

Euroopan unioni. *Yleinen tietosuoja-asetus.* Saatavilla WWW-muodossa, https://europa.eu/youreurope/business/dealing-with-customers/data-protection/data-protection-gdpr/index_fi.htm, viitattu 25.4.2019.

Fan, W., H. Wang, P. S. Yu ja S. Ma. 2003. “Is random model better? On its accuracy and efficiency”. Teoksessa *Third IEEE International Conference on Data Mining,* 51–58. doi:10.1109/ICDM.2003.1250902.

Fong, Pui Kuen, ja J. H. Weber-Jahnke. 2012. “Privacy Preserving Decision Tree Learning Using Unrealized Data Sets”. *IEEE Transactions on Knowledge and Data Engineering* 24 (2): 353–364. doi:10.1109/TKDE.2010.226. https://ieeexplore.ieee.org/document/5620916.

Han, Jiawei, Micheline Kamber ja Jian Pei. 2011. *Data Mining: Concepts and Techniques.* Saint Louis: Elsevier Science & Technology. ISBN: 9780123814807.

Huang, Zhengli, Wenliang Du ja Biao Chen. 2005. “Deriving Private Information from Randomized Data”. Teoksessa *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data,* 37–48. SIGMOD ’05. Baltimore, Maryland: ACM. ISBN: 1-59593-060-4. doi:10.1145/1066157.1066163.

Kargupta, H., S. Datta, Q. Wang ja Krishnamoorthy Sivakumar. 2003. “On the privacy preserving properties of random data perturbation techniques”. Teoksessa *Third IEEE International Conference on Data Mining,* 99–106. doi:10.1109/ICDM.2003.1250908.

Lindell, Yehuda, ja Benny Pinkas. 2000. “Privacy Preserving Data Mining”. Teoksessa *Advances in Cryptology — CRYPTO 2000,* toimittanut Mihir Bellare, 36–54. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-540-44598-2.

Malik, M. B., M. A. Ghazi ja R. Ali. 2012. “Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects”. Teoksessa *2012 Third International Conference on Computer and Communication Technology,* 26–32. doi:10.1109/ICCCT.2012.15.

Mendes, R., ja J. P. Vilela. 2017. “Privacy-Preserving Data Mining: Methods, Metrics, and Applications”. *IEEE Access* 5:10562–10582. doi:10.1109/ACCESS.2017.2706947.

New York Times. 2018. *Facebook Security Breach Exposes Accounts of 50 Million Users.* Saatavilla WWW-muodossa, https://nyti.ms/2OmvHom, viitattu 29.3.2019.

Quinlan, J. R. 1986. “Induction of decision trees”. *Machine Learning* 1 (1): 81–106. ISSN: 1573-0565. doi:10.1007/BF00116251. https://doi.org/10.1007/BF00116251.

Senosi, A., ja G. Sibiya. 2017. “Classification and evaluation of Privacy Preserving Data Mining: A review”. Teoksessa *2017 IEEE AFRICON,* 849–855. doi:10.1109/AFRICON.2017.8095593.

Vaidya, J., B. Shafiq, W. Fan, D. Mehmood ja D. Lorenzi. 2014. “A Random Decision Tree Framework for Privacy-Preserving Data Mining”. *IEEE Transactions on Dependable and Secure Computing* 11 (5): 399–411. doi:10.1109/TDSC.2013.43.

Vaidya, Jaideep, ja Chris Clifton. 2005. “Privacy-Preserving Decision Trees over Vertically Partitioned Data”, 3654:139–152. Data and Applications Security XIX. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 354028138X. doi:10.1007/11535706_11.

Xu, Lei, Chunxiao Jiang, Jian Wang, Jian Yuan ja Yong Ren. 2014. “Information Security in Big Data: Privacy and Data Mining”. *IEEE Access* 2:1149–1176. doi:10.1109/ACCESS.2014.2362522. https://ieeexplore.ieee.org/document/6919256.