Tampereen yliopisto

Eemeli Pesonen

KONTAMINAATION HAVAITSEMINEN TALOUSVESIVERKOSTOISSA

Tekniikan ja luonnontieteiden tiedekunta
Kandidaatintyö
Toukokuu 2019

TIIVISTELMÄ

Eemeli Pesonen: Kontaminaation havaitseminen talousvesiverkostoissa
Detection of a contamination in drinking water distribution systems
Kandidaatintyö
Tampereen yliopisto
Tekniikan ja luonnontieteiden TkK-tutkinto-ohjelma, Ympäristö- ja energiatekniikka
Toukokuu 2019

Varhaisvaroitusjärjestelmät (EWS) ovat järjestelmiä, jotka pyrkivät talous- tai raakavettä monitoroimalla havaitsemaan veteen päässeitä vieraita aineita. EWS:n tehtävänä on seurata veden laatuparametrejä erilaisilla sensoreilla, tulkita sensorien mittaamaa dataa ja auttaa päätöksenteossa ja kommunikoinnissa kontaminaation tapahtuessa. Tapahtumanhavaitsemisjärjestelmä (EDS) on EWS:n osa, joka tulkitsee reaaliajassa talousveden laatuparametrejä ja pyrkii havaitsemaan vedestä vieraiden aineiden aiheuttamat kontaminaatiot. EDS:t voivat hyödyntää esimerkiksi erilaisia koneoppimisalgoritmeja, jotka havaitsevat laatuparametrien normaalista poikkeavan käytöksen kontaminaation tapahtuessa. EDS:n tulisi havaita kontaminaatio mahdollisimman suurella todennäköisyydellä samalla, kun todennäköisyys väärälle hälytykselle on mahdollisimman pieni. Tässä työssä tutkitaan kolmea eri EDS-sovellutusta ja verrataan niiden suorituskykyjä keskenään. Työ on kirjallisuuskatsaus ja EDS:ien suorituskyvyt on selvitetty tutkimusartikkeleista. Työssä pohditaan sopivatko EDS:t sellaisinaan talousveden monitorointiin, vai tarvitaanko vielä lisää tutkimusta esimerkiksi erilaisilla kontaminaatioilla ja eri tavalla optimoiduilla algoritmeilla.

Tutkittavat EDS:t ovat painotettu tukivektorikone (SVM), pienimmän tilavuuden ellipsoidi -lajittelumalli (MVE) ja kanoninen korrelaatioanalyysi (CCA). EDS:istä SVM ja MVE ovat koneoppimisalgoritmeja. SVM ja MVE eroavat toisistaan muun muassa siten, että SVM:n harjoitus tapahtuu ohjatusti, kun taas MVE:n harjoitus tapahtuu ohjaamattomasti. CCA on tilastollinen menetelmä, jossa uusien mittaustulosten korrelaatiota verrataan vanhaan dataan. SVM:n ja MVE:n testaamiseen käytetty data on saatu oikean talousvesivarkoston normaalista ajosta, johon on lisätty keinotekoisia sattumanvaraisia kontaminaatioita. CCA:n testaamiseen on käytetty dataa laboratoriokokeesta, jossa normaaliin veteen on lisätty vieraana aineen akryyliamidia.

EDS:ien suorituskykyjen perusteella paras EDS on CCA. CCA ei aiheuttanut testauksessa ainuttakaan väärää hälytystä ja sen todennäköisyys havaita kontaminaatio vaihteli välillä 0,990 – 0,998. MVE:n todennäköisyys havaita kontaminaatio vaihteli välillä 0,61 – 1,00, ja mikäli MVE:n vääriä hälytyksiä saadaan karsittua, se voi olla varteenotettava vaihtoehto CCA:lle. EDS:istä huonoimmat suorituskyvyn arvot sai SVM. SVM:llä todennäköisyys kontaminaation havaitsemiseen oli välillä 0,44 – 0,98 ja vääriä hälytyksiä tuli enemmän kuin CCA:lla ja MVE:llä. Vaikka CCA vaikuttaa suorituskykyjen perusteella parhaalta EDS:ltä, CCA:n tuloksia on verrattava muihin EDS:iin varautuen, sillä CCA:n testaukseen käytettiin huomattavasti vähemmän dataan kuin MVE:n ja SVM:n testaukseen. Lisäksi CCA:n testauksen normaali vesi oli raakavettä talousveden sijaan. Tämä tarkoittaa sitä, että erityisesti CCA tarvitsee vielä jatkotutkimusta, jos menetelmää aiotaan käyttää talousveden monitorointiin. Jatkotutkimuksissa voisi tutkia muun muassa CCA:n suorituskykyä erilaisella puhtaalla vedellä ja eri lisättävillä vierailla aineilla. Tulosten perusteella SVM:ää ei todennäköisesti kannata tutkia enempää.

Avainsanat: talousvesi, EDS, EWS, koneoppiminen, online-mittaus, SVM, MVE, CCA, kontaminaatio

SISÄLLYSLUETTELO

1. JOHDANTO 1
2. TAPAHTUMANHAVAITSEMISJÄRJESTELMÄT (EDS) 3
2.1 Tapahtumanhavaitsemisjärjestelmän suorituskyky 4
3. KONEOPPIMISEN PERUSTEET 7
3.1 Koneoppimismenetelmien luokittelu 7
3.2 Harjoitusvaiheessa huomioitavaa 8
3.2.1 Harjoitusdatan määrä ja laatu 9
3.2.2 Algoritmit 10
3.2.3 Mallin testaus ja validointi ennen käyttöönottoa 11
4. EDS-SOVELLUTUSTEN VERTAILU 12
4.1.1 Painotettu tukivektorikone (SVM) 12
4.1.2 Pienimmän tilavuuden ellipsoidi -lajittelumalli (MVE) 13
4.1.3 Kanoninen korrelaatioanalyysi (CCA) 14
4.2 Algoritmit käytännössä 14
4.2.1 Testaamiseen käytetty data 14
4.2.2 Saadut tulokset 16
4.3 Pohdintaa tuloksista 17
4.4 Pohdintaa EDS:ien soveltuvuudesta talousvesiverkoston valvontaan 19
5. YHTEENVETO JA JOHTOPÄÄTÖKSET 20
LÄHTEET 22

LYHENTEET JA MERKINNÄT

CCA	canonical correlation analysis (kanoninen korrelaatioanalyysi)
EWS	early warning system (varhaisvaroitusjärjestelmä)
EDS	event detection system (tapahtumanhavaitsemisjärjestelmä)
FAR	false alarm rate (todennäköisyys antaa väärä hälytys)
FN	false negative (väärä negatiivinen tulos)
FP	false positive (väärä positiivinen tulos)
MVE	minimum volume ellipsoid (pienimmän tilavuuden ellipsoidi)
PD	probability of detection (tapahtumanhavaitsemistodennäköisyys)
SVM	support vector machine (tukivektorikone)
TN	true negative (oikea negatiivinen tulos)
TOC	total organic carbon (orgaaninen kokonaishiili)
TP	true positive (oikea positiivinen tulos)

1. JOHDANTO

Talousvesijärjestelmät ovat alttiita erilaisille kontaminaatioille, ja kontaminaation tapahtuessa kyseiseen järjestelmään kuuluvilla asiakkailla on riski saada terveydellisiä haittavaikutuksia käyttämästään vedestä. Terveydelle haitallisia aineita voi päästä jakeluverkostoon joko vahingossa, kuten Nokian vesikriisissä 2007 (Onnettomuustutkintakeskus, 2007), luonnollisen ilmiön seurauksena, kuten rikkoutuneen putken kautta, tai niitä voidaan johtaa verkostoon tahallisesti terroristisissa tarkoituksissa. Koska talousvesi kulkee yleensä verrattain nopeasti vesilaitokselta jakelupisteeseen ja leviää laajalle alueelle, on tärkeää, että kontaminaation tapahtuessa siihen pystytään reagoimaan nopeasti ja siitä aiheutuvat haitat pystytään minimoimaan. Tätä varten talousvesijärjestelmille on kehitetty varhaisvaroitusjärjestelmiä (EWS, early warning system), joiden tarkoitus on havaita ajoissa mahdollinen kontaminaatio talousvettä monitoroimalla ja kontaminaation tapahtuessa auttaa jatkotoimenpiteisiin liittyvässä päätöksenteossa sekä helpottaa kommunikointia viranomaisten kanssa (Hasan, 2005).

Jokaisen yksittäisen terveydelle haitallisen aineen monitorointi talousvedessä olisi joko taloudellisesti kannattamatonta, tai kokonaan mahdotonta. Tästä syystä talousvedestä täytyy jollain tapaa pystyä tunnistamaan kontaminaatiot yleisemmällä tasolla, vaikka perinteisiä veden laatuparametrejä tarkastelemalla. Varhaisvaroitusjärjestelmän tällaista osaa, joka pyrkii tunnistamaan talousveden kontaminaatiot tutkimalla talousveden laatuparametrejä, kutsutaan tapahtumanhavaitsemisjärjestelmäksi (EDS, event detection system). EDS tulkitsee reaaliajassa EWS:n sensorien mittaamaa dataa ja varoittaa havaitessaan normaalista ajosta poikkeavaa dataa. (U.S. EPA, 2010) Jos EDS on kalibroitu oikein niin, että se havaitsee luotettavasti talousvesiverkostoon päässeet vieraat aineet eikä se aiheuta vääriä hälytyksiä, se suojelee talousvesijärjestelmän asiakkaiden terveyttä ja vähentää kontaminaation tapahtuessa syntyvää taloudellista ja vahinkoa.

Tässä kandidaatin työssä tarkastellaan kolmen erilaista EDS:ää. Työn tavoitteena on selvittää, soveltuvatko tarkasteltavat EDS:t sellaisenaan talousveden monitorointiin ja mikä tarkasteltavista EDS:istä on tällä hetkellä suorituskyvyltään paras. Työ on kirjallisuuskatsaus, ja siinä verrataan keskenään tutkimuksissa saatuja tuloksia eri tapahtumanhavaitsemisalgoritmien toiminnasta ja kyvystä havaita kontaminaatio talousvesiverkostosta. Vertailtavia ominaisuuksia ovat muun muassa algoritmin kyky havaita oikeat kontaminaatiot ja välttää väärät hälytykset. Työssä käydään läpi myös tarvetta ja kannattavuutta jatkotutkimuksille kyseisten EDS:ien osalta. Tavoitteena on myös selvittää, mitä talousvesilaitos joutuu huomioimaan vertaillessaan erilaisia EDS:iä.

Työn toisessa luvussa esitellään, mikä on EDS sekä käydään läpi siihen liittyviä käsitteitä. Toisessa luvussa kerrotaan esimerkiksi, mitkä ominaisuudet tekevät EDS:stä luotettavan ja mikä rooli talousvedestä monitoroitavilla laatuparametreillä on. Kolmannessa luvussa käydään läpi

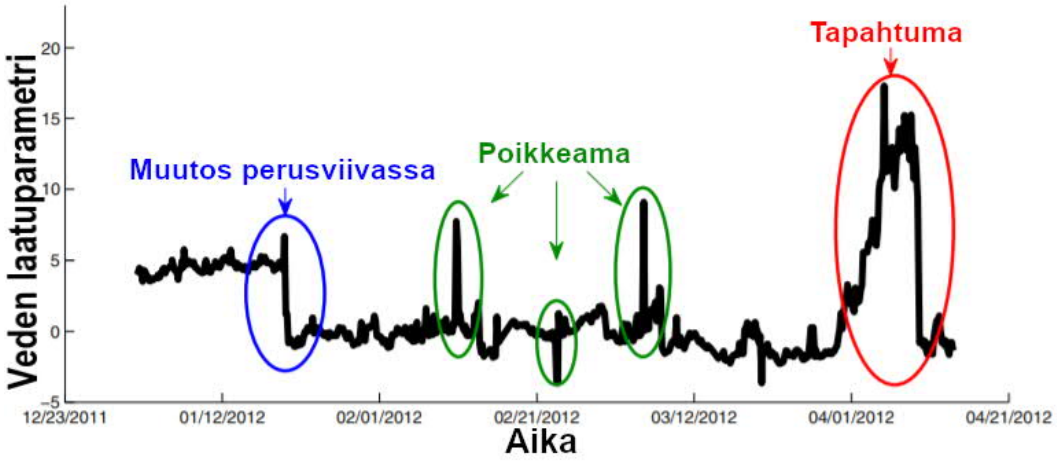
yleisellä tasolla koneoppimisen perusteet. Luvussa esitellään erilaisia tapoja opettaa algoritmi ja kerrotaan muun muassa käytettävän datan laatuvaatimuksista. Kolmannen luvun lopussa kerrotaan ongelmista, joita saattaa esiintyä harjoitusdatalla luodussa koneoppimismallissa, ja siitä miten nämä ongelmat voidaan havaita testaamalla malli. Neljännessä luvussa esitetään erilaisia tutkittuja koneoppimisalgoritmeja ja niillä saatuja tuloksia talousveden testidatan monitoroinnissa. Luvussa myös vertaillaan tuloksia keskenään, ja mietitään eri algoritmien vahvuuksia ja heikkouksia. Viidennessä luvussa on yhteenveto ja johtopäätökset kandidaatintyöstä.

2. TAPAHTUMANHAVAITSEMISJÄRJESTELMÄT (EDS)

Tapahtumanhavaitsemisjärjestelmät tulkitsevat reaaliajassa talousvesiverkoston sensorien mittaamaa dataa veden laatuparametreistä ja pyrkivät erottamaan vedestä kontaminaation aiheuttamat muutokset verkoston normaalin ajon aikaisista laadun muutoksista. Nämä normaalit muutokset voivat johtua muun muassa vuorokaudenajasta, vuodenajasta tai erilaisista talousvesiverkostossa tapahtuvista operaatioista. Kontaminaation aiheuttaman muutoksen toteaminen tapahtuu yleensä tulkitsemalla niin kutsuttuja korvikeparametrejä (surrogate parameters). Korvikeparametrit ovat parametrejä, jotka eivät suoraan kuvaa tietyn saasteen määrää vedessä. EDS voi kuitenkin korvikeparametrien poikkeavaa käyttäytymistä tutkimalla päätellä onko kyseessä kontaminaatio. Tämä tarkoittaa sitä, että talousvesiverkostossa ei tarvitse monitoroida jokaista mahdollista terveydelle haitallisen aineen pitoisuutta erikseen, vaan riittää että EDS seuraa talousveden perinteisiä laatuparametrejä ja havaitsee niistä kontaminaation aiheuttamat muutokset. (U.S.EPA, 2010)

EDS:ien korvikeparametreinä voidaan käyttää perinteisiä veden laatuparametrejä, joita vesilaitokset monitoroivat monesti muutenkin. Useissa EDS-tutkimuksissa korvikeparametreinä on käytetty muun muassa pH:ta, sameutta, UV-valoa, sähkönjohtavuutta, TOC:ia, lämpötilaa sekä verkoston vapaata klooria (Hasan, 2005; Whalen et al., 2010; Murray et al., 2011). Edellä mainitut parametrit sopivat hyvin EDS:lle, sillä niitä mittaavat sensorit ovat luotettavia ja vaativat vähäistä huoltoa (U.S.EPA 2005). EDS:issä toimivat korvikeparametrit eivät kuitenkaan rajoitu edellä mainittuihin parametreihin ja esimerkiksi Whalen et al. artikkelissa on käytetty vedestä mitattavan adenosiinitrifosfaatin (ATP) määrää todetun kontaminaation varmistamiseksi (Whalen et al. 2010).

Kuva 1 (Zhao et al., 2014) esittää tilannetta, jossa veden laatuparametrin mittausdatasta on piirretty kuvaaja. Kuvan tapauksessa varhaisvaroitusjärjestelmässä sensori tuottaa reaaliajassa EDS:lle dataa, jossa näkyy perusviiva (baseline), joka kuvaajalle piirrettynä näyttää melko tasaiselta viivalta tai joltain toistuvalta kuviolta. Esimerkiksi vuorokaudenajan vaihtelu voi aiheuttaa normaaliksi luokiteltavaa toistuvaa heilahtelua kuvaajan perusviivassa. Kuvan sinisellä värillä merkatussa kohdassa tapahtuu pysyvä muutos perusviivassa. Muutos voi johtua esimerkiksi tilanteesta, jossa normaalisti pohjavettä jakavaan talousvesiverkostoon joudutaan johtamaan vettä, joka on peräisin pintavedesistä. Vihreällä värillä merkityt kuvan 1 piikit kuvaavat poikkeamia (outlier) perusviivasta. Mikäli poikkeamat näkyvät datassa vain lyhyen ajan, ei yleensä ole kyseessä kontaminaatio. Jos poikkeama kuitenkin jatkuu pidempään tai niitä on useita peräkkäin, puhutaan tapahtumasta (event), joka EDS:n täytyy tunnistaa ja aiheuttaa hälytys. EDS:n pitää siis osata erottaa varsinaiset tapahtumat poikkeamista eikä se saa aiheuttaa turhia hälytyksiä, kun perusviivassa tapahtuu lyhyt luonnollinen muutos.

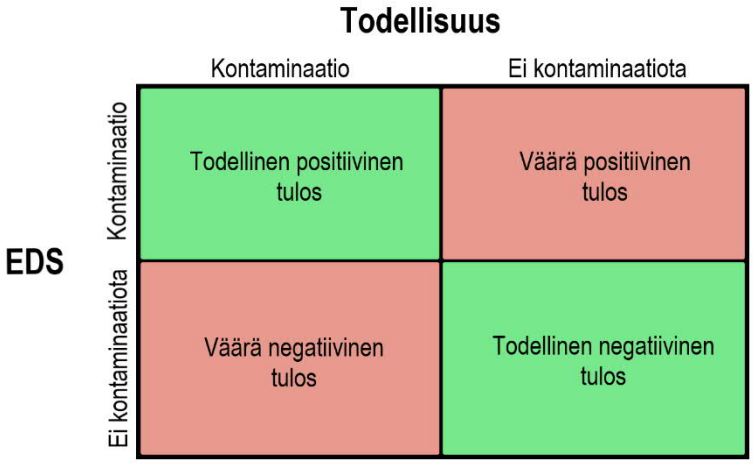


Kuva 1. *Esimerkki mittausdatassa esiintyvästä muutoksesta perusviivassa, poikkeamasta sekä tapahtumasta. (Muokattu lähteestä Zhao et al., 2014)*

Modernit EDS:t käyttävät kontaminaation toteamiseen useaa korvikeparametriä. Seuraamalla useaa parametriä saadaan luotettavampia tuloksia kontaminaation havaitsemisessa, sekä voidaan mahdollisesti tunnistaa kontaminaation aiheuttaja. (Oliker & Ostfeld, 2014a; Oliker & Ostfeld 2014b; Li et al., 2016) Tutkimuksissa on todettu, että vieraat aineet voivat aiheuttaa samanaikaisesti usealle veden laatuparametrille tunnistettavia poikkeamia vertailupisteestä (Byer et al., 2005; Hall et al., 2007). Eri tapahtumanhavaitsemisalgoritmeilla on erilaiset menetelmät määrittää poikkeamaksi laskettavat heilahtelut perusviivasta.

2.1 Tapahtumanhavaitsemisjärjestelmän suorituskyky

EDS voi tulkita mittaustuloksia väärin kahdella eri tavalla. Kun EDS tulkitsee normaalin ajon aikaisen mittaustuloksen tapahtumana, kyseessä on väärä positiivinen tulos (false positive, FP). Jos taas EDS ei havaitse todellista tapahtumaa ja väittää kyseessä olevan normaali mittaustulos, puhutaan väärästä negatiivisesta tuloksesta (false negative, FN). Väärien positiivisten ja negatiivisten tulosten lisäksi puhutaan oikeista positiivisista tuloksista (true positive, TP) ja oikeista negatiivisista tuloksista (true negative, TN). Oikeassa positiivisessa tuloksessa EDS on huomannut todellisen kontaminaation ja oikeissa negatiivisissa tuloksissa ei ole tapahtunut vääriä hälytyksiä. (Spence et al., 2013) Kuvassa 2 on visualisoitu yllä mainitut neljä tapausta. EDS:n luotettavuutta ja suorituskykyä määriteltäessä on tarkasteltava EDS:n väärien positiivisten tulosten ja väärien negatiivisten tulosten määrää suhteessa kaikkiin tuloksiin.



Kuva 2. *Todellinen tilanne ja tapahtumanhavaitsemisjärjestelmän (EDS) tulkitsema tilanne monitoroitavasta mittausdatasta. Vaakasuora akseli kuvaa todellista tilannetta ja pystysuora akseli kuvaa tapahtumanhavaitsemisjärjestelmän tulkintaa tilanteesta. (Muokattu lähteestä Spence et al., 2013)*

Suorituskyvyltään hyvä EDS ei tuota kuvan 2 punaisella merkittyjä vääriä positiivisia ja vääriä negatiivisia tuloksia. Jos EDS aiheuttaa hälytyksen väärän positiivisen tuloksen takia, on seurauksena todennäköisesti kustannuksia ja turhaa työtä vesilaitokselle (tai muulle veden valvonnasta vastaavalle taholle) sekä viranomaisille, joille tieto kontaminaatiosta välitetään. Tämän lisäksi vesilaitosten asiakkaiden luottamus hälytysten vakavuuteen pienenee, mikäli vääriä hälytyksiä tapahtuu usein. Väärä negatiivinen tulos taas aiheuttaa mahdollisen terveyshaitan talousveden käyttäjille, sillä varoitus kontaminaatiosta talousvedessä jää tekemättä. (Dejus et al., 2017)

EDS:n testausvaiheessa havaituista vääristä positiivisista ja negatiivisista tuloksista voidaan johtaa laskukaavoja EDS-menetelmien vertailuun ja niiden toiminnan arvioimiseen. Usein käytetty suure kuvaamaan EDS:n kykyä havaita todelliset positiiviset tapahtumat, on tapahtumanhavaitsemistodennäköisyys (propability of detection), jota merkitään kaavoissa lyhenteellä PD. PD on EDS:n oikein tulkitsemien tapahtumien suhde kaikkiin tapahtumiin (kaava 1). Olikerin ja Ostfeldin artikkeleissa (Oliker & Ostfeld, 2014a; Oliker & Ostfeld, 2014b) on käytetty PD:n sijasta termiä ”detection ratio”, joka kuitenkin lasketaan samalla tavalla (kaava 1). PD lasketaan seuraavasti:

$$PD = \frac{TP}{TP+FN}, \tag{1}$$

jossa TP on todellisten positiivisten tulosten määrä ja FN on väärien negatiivisten tulosten määrä. PD:n lisäksi EDS:n suorituskykyä voidaan mitata EDS:n todennäköisyydellä tuottaa väärä hälytys (false alarm rate, FAR). FAR kertoo kuinka todennäköisesti EDS tekee vääriä hälytyksiä normaalin toiminnan aikana (kaava 2). (Liu et al., 2015a)

$$FAR = \frac{FP}{FP+TN} \tag{2}$$

Tukivektorikoneita ja pienimmän tilavuuden ellipsoideja käsittelevissä artikkeleissa (Oliker & Ostfeld, 2014a; Oliker & Ostfeld, 2014b) yhtenä suorituskyvyn mittarina on käytetty oikeellisuutta (accuracy), joka on EDS:n oikein tulkitsemien tulosten suhde kaikkiin tuloksiin (kaava 3).

$$oikeellisuus = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

Kaikki tässä luvussa mainitut suorituskyvyn mittaussuuret pysyvät välillä 0 – 1. Kaavoista (1), (2) ja (3) nähdään, että hyvä EDS on sellainen, jolla PD ja oikeellisuus ovat mahdollisimman suuria ja FAR:n tulisi olla mahdollisimman pieni.

3. KONEOPPIMISEN PERUSTEET

Koneoppimisen perusideana on, että algoritmi ensin opetetaan tekemään haluttu tehtävä kouluttamalla se harjoitusvaiheessa harjoitusjoukon (training set) avulla, minkä jälkeen algoritmi osaa suorittaa opitun tehtävän sille täysin uudella datalla. Algoritmi luo harjoitusdatalla itselleen niin kutsutun mallin (model), jota se sitten käyttää kohdatessaan uusia ongelmia. Koneelle ei tarvitse siis ohjelmoida manuaalisesti toimintaa jokaista mahdollista tilannetta varten. (Louridas & Ebert, 2016) Koneoppiminen sopii hyvin kontaminaation tunnistamiseen talousvesiverkosta, sillä verkostojen toiminta on erittäin monimutkaista ja monitoroitavien laatuparametrien käyttäytymistä vedessä ei tunneta täydellisesti. Koneoppimisen avulla mittausdatasta voidaan havaita kontaminaatio ilman, että korvikeparametrien reagoimista epäpuhtauden kanssa tunnetaan täydellisesti. (Mounce et al., 2014) Lisäksi jos koneoppimismalli saadaan toimimaan hyvin, voidaan mallia ja sen oppimista tutkimalla saada kokonaan uutta tietoa kohteesta, johon koneoppimismallia on sovellettu. (Géron, 2017). Mallia tutkimalla voi esimerkiksi paljastua uusia mielenkiintoisia korrelaatioita vieraiden aineiden ja veden perinteisten laatuparametrien välillä. Tässä luvussa käydään lyhyesti läpi koneoppimisen perusteita ja koneoppimismenetelmien jaottelua muun muassa erilaisten koulutustapojen perusteella.

3.1 Koneoppimismenetelmien luokittelu

Koneoppimisessa varsinainen harjoitusvaihe voidaan jaotella Louridaksen ja Ebertin mukaan kahteen oppimisstrategiaan: ohjaamattomaan oppimiseen (unsupervised learning) ja ohjattuun oppimiseen (supervised learning). Ohjaamattomassa oppimisessa tietokoneelle annetaan harjoitusjoukko, joka sisältää harjoitusdatan, mutta koneelle ei anneta oikeita vastauksia eli ulostuloja (output). Tietokoneen täytyy löytää annetusta harjoitusdatasta säännöllisyydet ja kehittää jonkinlainen malli säännöllisyyksien muodostumiselle. (Alpaydin, 2014; Louridas & Ebert, 2016) Yksi tällainen ohjaamattoman oppimisen alalaji on klusterointi (clustering), jossa kone pyrkii muodostamaan datasta ryhmiä datapisteiden samankaltaisuuksien perusteella (Alpaydin, 2014). Talousvesijärjestelmän valvonnan kannalta ohjaamaton oppiminen on kiinnostava koneoppimisen osa-alue, sillä yksi ohjaamattoman oppimisen alaisuuteen luokiteltava tehtävä on poikkeamien erottaminen (anomaly detection) normaalista datasta. Poikkeamien erotuksessa malli koulutetaan normaalilla datalla, minkä jälkeen se osaa kertoa, mikäli uusi data poikkeaa normaalista tilanteesta. (Géron, 2017) EDS:issä pyritään siis siihen, että luotu koneoppimismalli käsittelee kontaminaatiot poikkeamina.

Ohjatussa oppimisessa algoritmille syötetään harjoitusjoukko, johon kuuluu syötteenä data sekä oikeat vastaukset kyseiselle datalle. Algoritmi oppii annetun harjoitusjoukon avulla, miten oikeat vastaukset saadaan aikaiseksi ja luo sen pohjalta itselleen mallin, jota voi käyttää kokonaan uuteen dataan. (Louridas & Ebert, 2016)

Ohjatun ja ohjaamattoman oppimisen lisäksi yksi koneoppimisen laji on niin kutsuttu vahvistusoppiminen (reinforcement learning). Vahvistusoppimisessa yksittäinen output ei ole tärkeä, vaan haluttuun lopputulokseen pyritään pääsemään sarjalla oikeanlaisia ulostuloja. Tässä oppimismallissa algoritmi tarvitsee palautetta (positiivista tai negatiivista) ulostulojen sarjoista, jolloin se voi kehittää käyttäytymistavan (policy), jonka perusteella ohjelma toimii jatkossa. Algoritmi pyrkii luomaan käyttäytymistavan, joka johtaa mahdollisimman suureen määrään positiivista palautetta. (Alpaydin, 2014)

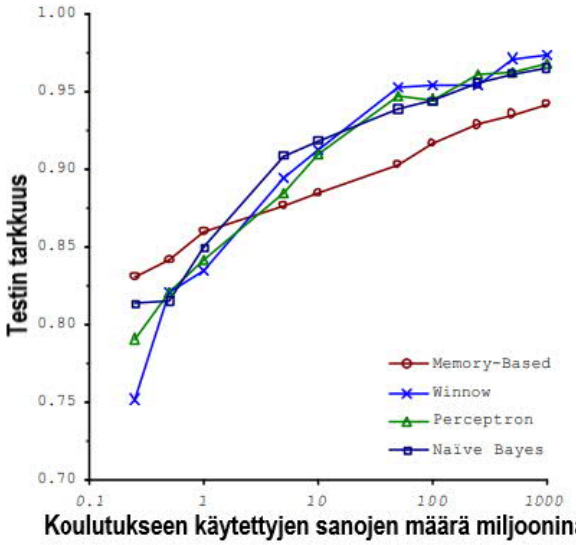
Koneoppimisalgoritmit on myös mahdollista jaotella Géronin mukaan oppimismenetelmän lisäksi sen perusteella, osaavatko algoritmit oppia reaaliajassa online-ympäristössä, vai tapahtuuko oppiminen offlinessa. Offlinessa tapahtuva oppiminen tarkoitta sitä, että ohjelma on koulutettava ennen käyttöönottoa, jonka jälkeen sitä ei voi enää käytön aikana kouluttaa uudella datalla. Mikäli ohjelmalle halutaan esittää uutta dataa, on koko koulutusvaihe tehtävä uudestaan sekä vanhalla että uudella koulutusdatalla, minkä jälkeen uusi versio ohjelmasta otetaan käyttöön. Online-oppivissa ohjelmissa on offline-oppiviin ohjelmiin verrattuna paremman päivitettävyyden lisäksi hyötynä se, että oppimismenetelmä vaatii vähemmän laskentatehoa ja tallennustilaa tietokoneelta. Jos ohjelmaa opetetaan isolla määrällä dataa ja käytössä oleva tallennustila on rajallinen, data voidaan online-ympäristössä oppivan ohjelman tapauksessa pilkkoa pienempiin osiin ja ohjelma voidaan opettaa osa kerrallaan. Online-oppivat ohjelmat eivät tarvitse harjoitusdataa harjoitusvaiheen jälkeen, joten datan voi poistaa, kun sitä ei enää tarvitse. (Géron, 2017)

3.2 Harjoitusvaiheessa huomioitavaa

Jotta koneoppimisalgoritmi toimisi mahdollisimman hyvin ja halutulla tavalla, on ohjelmaa testattaessa syytä varmistaa, että harjoitusvaiheessa käytettävä data on todellista tilannetta kuvaavaa, jotta sen pohjalta saadaan ohjelmalle toimiva malli. Harjoitusvaiheessa syntynyttä mallia täytyy myös tutkia sen varalta, että se ei esimerkiksi anna liikaa painoarvoa koulutusdatassa esiintyvälle kohinalle. Lopuksi ennen ohjelman käyttöönottoa ohjelma ja sen toiminta täytyy testata mallille uudella datalla. (Géron, 2017)

3.2.1 Harjoitusdatan määrä ja laatu

Toimiakseen hyvin koneoppimisalgoritmit tarvitsevat tarpeeksi suuren määrän dataa koulutusvaiheeseen (Géron, 2017). Tästä esimerkkinä toimii Bankon ja Brillin 2001-vuonna tekemä tutkimus (Banko & Brill, 2001), jossa tutkittiin, miten harjoitusdatan määrä vaikuttaa koneoppimisalgoritmin kykyyn tunnistaa algoritmille syötetyn tekstin kieli (kuva 3).



Kuva 3. *Koulutukseen käytettyjen sanojen määrän vaikutus koneoppimismallin tarkkuuteen. Kuvassa eri värillä ja eri muodoilla tehdyt kuvaajat kuvaavat eri algoritmien suorituskykyä. (Muokattu lähteestä Banko & Brill, 2001)*

Tutkimus osoitti, että jos dataa on saatavilla tarpeeksi, itse algoritmin monimutkaisuudella tai kehittyneisyydellä ei välttämättä ole merkitystä (Banko & Brill, 2001). Kuitenkaan aina ei ole mahdollista saada ideaaleja määriä harjoitusdataa. Voi olla, että esimerkiksi vesilaitoksella ei ole tarpeeksi historiallista dataa laitoksen normaalista ajosta, jolloin itse koneoppimisalgoritmin valinta on tärkeää.

Jos koulutusdata ei edusta todellista tilannetta ja datan perusteella luotu malli antaa vääriä positiivisia ja negatiivisia tuloksia, käytettyä koulutusdataa kutsutaan epäedustavaksi (nonrepresentative data). Data voi olla epäedustavaa esimerkiksi silloin, kun koulutukseen on käytetty liian pientä määrä harjoitusdataa, milloin epäedustavuus johtuu usein sattumasta. Pienen dataotoksen aiheuttamaa epäedustavuutta kutsutaan näytteenottokohinaksi (sampling noise). (Géron, 2017) Oliker ja Ostfeld pyrkivät artikkeleissaan pienentämään talousvesiverkon näytteenottokohinaa käyttämällä tarpeeksi dataa (neljän viikon ajalta) ja poistamalla mittausparametreille mahdottomat mittaustulokset (esimerkiksi negatiiviset tulokset) datan esikäsittelyssä (Oliker & Ostfeld, 2014a; Oliker & Ostfeld, 2014b).

Suurikaan koulutusdatamäärä ei välttämättä ole edustavaa, jos data on kerätty puolueellisesti. Tällaisissa tapauksissa kyse on puolueellisuudesta (sampling bias). (Géron, 2017) Otetaan yksinkertaisena esimerkkinä kuvitteellinen tilanne, jossa tietokone päättelee koneoppimisalgoritmia käyttäen Suomen korkeakoulujen sukupuolijakauman. Algoritmi käyttää koulutusjoukkonaan vain teknillisten korkeakoulujen opiskelijoiden sukupuolijakaumaa. Oletetaan, että suurempi osa teknillisten korkeakoulujen opiskelijoista on miehiä, jolloin

tietokoneen mukaan tilanne on todennäköisesti sama myös kaikissa muissa Suomen korkeakouluissa, vaikka näin ei todellisuudessa olisikaan.

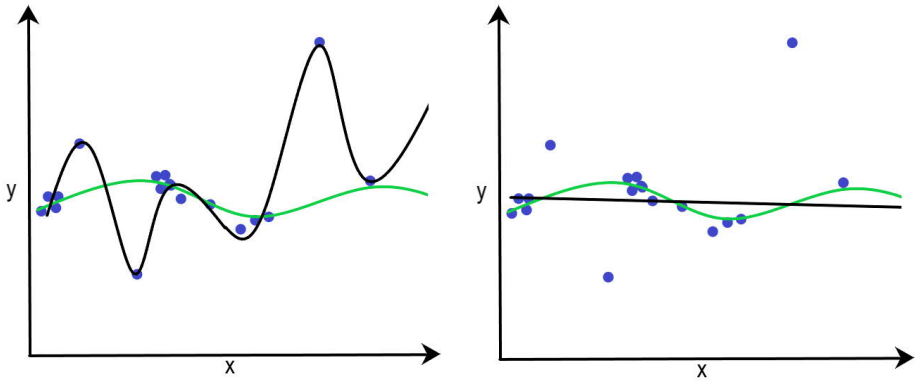
Koneoppimisalgoritmille ei kannata syöttää mitä tahansa dataa harjoitusvaiheessa. Jos harjoitusdatassa on paljon kohinaa, mittausvirheistä johtuvia poikkeamia tai esimerkiksi puuttuvia mittaustuloksia, se vaikuttaa negatiivisesti ohjelman toimintaan ja mallin luomiseen. Lisäksi turhien mittausparametrien syöttäminen algoritmille harjoitusvaiheessa voi aiheuttaa ohjelman laadun heikkenemistä. Joissain tapauksissa on siis kannattavaa esikäsitellä syötettävää dataa ja miettiä, mitkä mittausparametrit ovat oleellisia. (Géron, 2017)

3.2.2 Algoritmit

Koulutuksessa käytettävän datan laadun ja sen sopivan määrän lisäksi on tärkeää, että käytettävä algoritmi on sopivia. Tilannetta, jossa koneoppimisalgoritmin luoma malli kuvaa liian tarkasti koulutusjoukkoa ja siinä esiintyvää kohinaa, kutsutaan ylioppimiseksi (overfitting). Ylioppinut malli toimii hyvin koulutusdatan kanssa, mutta uutta dataa se ei osaa generalisoida. Ylioppimista voidaan ehkäistä esikäsittelemällä dataa ja miettimällä koulutuksessa käytettävien parametrien määrän rajoittamista. Vaihtoehtona on myös luodun mallin rajoittaminen. Tällaista mallin rajoittamista ja sen yksinkertaistamista kutsutaan regularisoinniksi (regularization) ja sen määrää voidaan säädellä muuttamalla ennen koulutusvaihetta koulutusalgoritmin niin kutsuttua hyperparametriä (hyper-parameter). Muokkaamalla hyperparametriä pienemmäksi saadaan malliin enemmän ylioppimista, ja vastaavasti suurentamalla hyperparametriä malli on tasaisempi ja yksinkertaisempi.

Vastakohta ylioppineelle mallille on liian tasainen malli. Tällaista mallia kutsutaan alioppineeksi. Hyperparametria hienosäätämällä voidaan tapauskohtaisesti tehdä koneoppimisohjelman mallista halutun kaltainen. (Géron, 2017)

Otetaan yli- ja alioppimisesta esimerkkinä mallit, joilla on kaksi parametriä x ja y, ja jotka voidaan esittää käyränä x-y-koordinaatistossa (kuva 4). Esimerkin harjoitusvaiheessa datassa on ollut jonkun verran kohinaa, ja mitattujen datapisteiden joukossa on muutama selkeä virheellinen mittaus. Ylioppineessa mallissa mallin käyrä mukailee tarkasti harjoitusdataa ja tuloksena on korkeamman asteen polynomi. Kuitenkaan harjoitusdatan kohinan takia uuden datan generalisointi ei tule vastaamaan todellisuutta, sillä harjoitusdatan kohinan takia luotu malli kuvittelee todellisen tilanteen olevan monimutkaisempi, kun se oikeasti on. Toisaalta alioppineessa mallissa mallin käyrä on suora ja malli on liian yksinkertainen kuvaamaan todellista tilannetta.



Kuva 4. *Vasemmalla esimerkki ylioppineesta mallista ja oikealla esimerkki alioppineesta mallista. Pisteet esittävät harjoitusdatapisteitä, vihreät vaaleat käyrät esittävät ideaalia mallia ja mustat käyrät esittävät algoritmin luomia malleja. (Muokattu lähteestä Géron, 2017)*

Esimerkin kaltaisissa ylioppineissa tilanteissa on todennäköisesti käytetty liian pientä hyperparametriä ja vastaavasti alioppineissa tapauksissa liian suurta. Esimerkin tapauksessa ideaali malli voitaisiin saavuttaa poistamalla selkeät virheelliset mittaukset datan esikäsittelyssä ja pitämällä hyperparametri kohtuullisen suurena.

3.2.3 Mallin testaus ja validointi ennen käyttöönottoa

Luotu koneoppimismalli ja sen toiminta on aina testattava ennen käyttöönottoa, jotta sen suorituskyky saadaan selville. Koneoppimisohjelman toiminta testataan usein jakamalla koulutuksessa käytettävä data harjoitusjoukoksi ja testijoukoksi (test set). Harjoitusdata annetaan ohjelmalle syötteenä, jonka avulla se luo itselleen mallin. Tämän jälkeen mallin toimintaa testataan testijoukolla, jota ohjelma ei ole ennen nähnyt. (Géron, 2017)

Talousvesiverkoston kontaminaation havaitsemisessa mallinluomisprosessi voisi mennä esimerkiksi seuraavalla tavalla: Ensin koneoppimisalgoritmi koulutetaan datalla, jossa veden laatuparametrit käyttäytyvät normaalilla tavalla. Koulutuksen jälkeen ohjelma testataan syöttämällä sille dataa, jossa on normaalin datan lisäksi simuloituja kontaminaatioita. Testin jälkeen katsotaan, monta oikeaa kontaminaatiota ohjelma havaitsi ja havaitsiko ohjelma testijoukosta olemattomia kontaminaatioita. Jos ohjelman PD ja oikeellisuus ei ole tarpeeksi hyvä, aloitetaan koulutus alusta ja mietitään esimerkiksi harjoitusdatan laadun ja käytetyn algoritmin vaikutusta huonoon lopputulokseen. Jos taas ohjelman toiminta on tyydyttävällä tasolla, se voidaan ottaa käyttöön.

4. EDS-SOVELLUTUSTEN VERTAILU

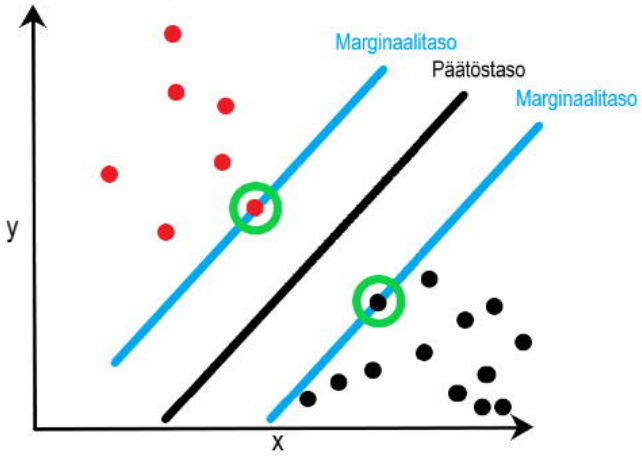
Tässä luvussa tarkastellaan kolmea eri tavalla toimivaa EDS:ää ja vertaillaan niiden kykyä havaita vieraat aineet talousvesiverkostosta. Aluksi esitetään lyhyesti käytettyjen algoritmien toimintaperiaate, jonka jälkeen vertaillaan näitä algoritmeja hyödyntämällä tutkimuksissa saatuja tuloksia.

Tässä kandidaatintyössä tarkasteltavia algoritmeja ovat painotettu tukivektorikone (weighted support vector machine, SVM), pienimmän tilavuuden ellipsoidi -lajittelumalli (minimum volume ellipsoid classification model, MVE) sekä kanoninen korrelaatioanalyysi (canonical correlation analysis, CCA). Tässä työssä tarkasteltavat EDS:t on valittu Dejus et al. artikkelissa esitettyjen suorituskykyjen perusteella (Dejus et al., 2017). Näistä kolmesta SVM ja MVE kuuluvat koneoppimisen piiriin ja CCA on tilastotieteellinen menetelmä (Oliker & Ostfeld, 2014a; Oliker & Ostfeld, 2014b; Li et al., 2016). Algoritmeista SVM:ää ja MVE:tä tutkitaan käytännössä Olikerin ja Ostfeldin artikkeleissa (Oliker & Ostfeld, 2014a; Oliker & Ostfeld, 20014b). Li et al. käsittelivät artikkelissaan CCA:ta (Li et al., 2016).

4.1.1 Painotettu tukivektorikone (SVM)

SVM:llä pystyy tulkitsemaan samanaikaisesti useaa mittausparametriä ja jaottelemaan ne kahteen eri joukkoon: esimerkiksi kontaminaatio tai ei kontaminaatiota. Kun SVM-malli luodaan, pyritään löytämään kahden joukon välistä sellainen taso (hyperplane), joka kulkee yhtä läheltä molempien joukkojen lähintä vektoria (yhtä datapistettä) (kuva 5). Näitä vektoreita kutsutaan tukivektoreiksi. Tukivektoreiden ”läpi” kulkee päätöstason suuntaiset marginaalitasot, joiden väliin ei jää harjoitusdatasta yksikään vektori. Ainoastaan tukivektorit vaikuttavat harjoitusvaiheessa luotavaan luokittelumalliin. (Boser et al., 1992; Cortes & Vapnik, 1995)

SVM:n koulutus tapahtuu ohjatusti, eli harjoitusvaiheessa algoritmille kerrotaan, onko harjoitusdatapisteiden kohdalla kyse kontaminaatiosta vai normaalista datasta. Harjoitusvaiheen jälkeen luotu malli tarkastaa kummalle puolelle päätöstasoa uudet vektorit sijoittuvat, ja sijoittumisen avulla päättää kumpaan luokkaan vektori kuuluu. (Oliker & Ostfeld, 2014a)

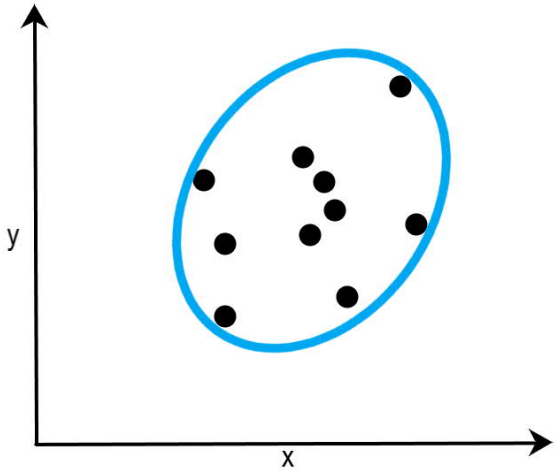


Kuva 5. *Tukivektorikoneen (SVM) perusperiaate. x- ja y-akselit kuvaavat mitattavia parametrejä. Tukivektorit näkyvät ympyröityinä pisteinä marginaalitasoilla. (Muokattu lähteestä Boser et al., 1992)*

Mikäli harjoitusdata ei ole täysin puolueetonta, tarvitaan painotettua SVM:ää. Esimerkiksi kun SVM-mallia opetetaan talousvesiverkoston datalla, datassa on vähemmän mittaustuloksia kontaminaatioista. Jotta luotu malli ei jaottele kaikkia uusia mittauksia normaaleiksi, mallia luodessa yhtälöt täytyy painottaa antamalla enemmän arvoa poikkeaville mittaustuloksille. (Oliker & Ostfeld, 2014a)

4.1.2 Pienimmän tilavuuden ellipsoidi -lajittelumalli (MVE)

MVE-lajittelumallin harjoitusvaiheessa luodaan ellipsoidi, johon kuuluu suurin osa harjoitusdatasta (esimerkiksi 95%) (kuva 6). Algoritmin harjoitus tapahtuu ohjaamattomasti eikä harjoitusdataan oteta mukaan poikkeamia. Täten esimerkiksi talousvesiverkoston EDS:n tapauksessa koulutuksessa ei käytetä dataa kontaminaatiotapauksista. Kuvassa 6 esitetään kaksiulotteinen esimerkki MVE-lajittelumallista. (Rousseeuw, 1985; Oliker & Ostfeld, 2014b)



Kuva 6. *Pienimmän tilavuuden ellipsoidi -lajittelumallin perusperiaate. x- ja y-akselit kuvaavat mitattavia parametrejä. Harjoitusdatan ympärille on luotu ellipsoidi, jonka sisälle harjoitusdatapisteet mahtuvat. (Muokattu lähteestä Oliker & Ostfeld, 2014b)*

Harjoitusvaiheen jälkeen MVE-malli tarkastaa osuvatko uudet mittaustulokset ellipsoidin sisälle. Jos näin ei ole, kyseessä on poikkeama. (Oliker & Ostfeld, 2014b)

4.1.3 Kanoninen korrelaatioanalyysi (CCA)

CCA:ssa tavoitteena on löytää kahden muuttujajoukon väliltä korrelaatioita ja määrittämään korrelaation vahvuus. Suhteen vahvuutta arvioidaan laskemalla niin sanottuja kanonisia korrelaatiokertoimia. (Härdle & Simar, 2007)

Li et al. artikkelissa (Li et al., 2016) CCA:ta käytettiin kontaminaation havaitsemiseen vertaamalla uusien mittaustulosten korrelaatiokertoimia kynnysarvoon. Kynnysarvo määritettiin kokeilemalla ja vertaamalla eri arvoilla saatuja suorituskyvyn arvoja keskenään. Suorituskyky testattiin yhteensä 1001 eri kynnysarvolla väliltä 0 – 1. Parasta suorituskykyä vastaava kynnysarvo valittiin ja tämä kynnysarvo otettiin käyttöön varsinaiseen suorituskyvyn testaamiseen. Vasta optimaalisella kynnysarvolla saatuja suorituskyvyn arvoja on järkevää käyttää algoritmin vertailemiseen muihin EDS:iin. (Li et al., 2016)

4.2 Algoritmit käytännössä

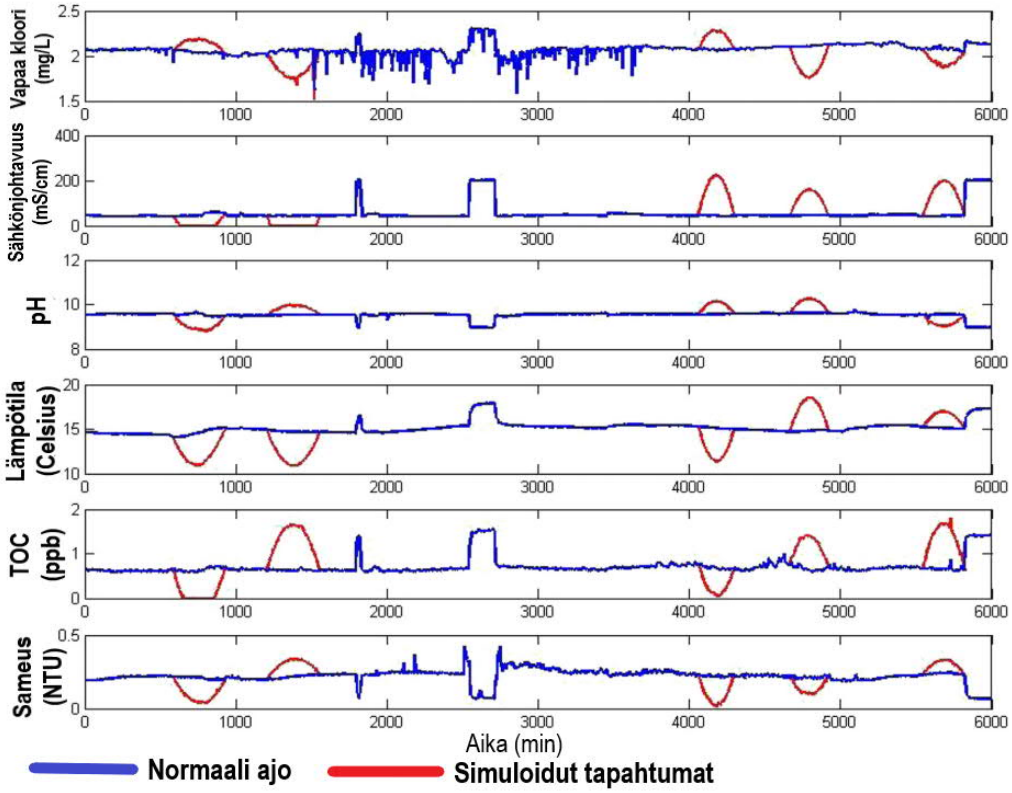
Luvussa 4.1 käsitellyt algoritmit havaitsevat poikkeamia, mutta on kuitenkin huomattava, että jokainen poikkeama ei ole tapahtuma. Vähentääkseen väärien hälytysten määrää, Oliker ja Ostfeld tarkastivat tutkimuksissaan vielä poikkeaman havaitsemisen jälkeen, onko aikasarjassa poikkeaman lähistöllä useita peräkkäisiä poikkeamia. EDS:ien suorituskyvyt voidaan arvioida tämän sekvenssianalyysin (sequence analysis) jälkeen. (Oliker & Ostfeld, 2014a; Oliker & Ostfeld, 20014b) Li et al. CCA-artikkelissa vastaavaa sekvenssianalyysiä ei käytetty (Li et al., 2016).

4.2.1 Testaamiseen käytetty data

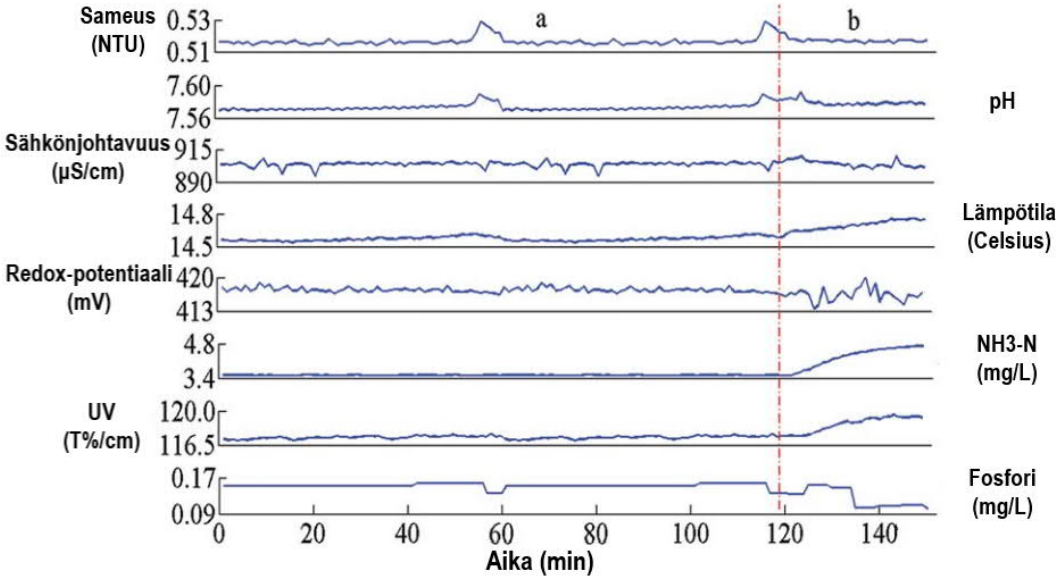
Olikerin ja Ostfeldin tutkimuksessa EDS:ien testaamiseen ja kouluttamiseen käytettiin dataa neljän viikon ajalta viiden minuutin välein mitattuna. Käytössä oli historiallista dataa oikeasta talousvesiverkosta, joka oli toiminut normaalisti koko tämän ajan. Tutkittavat veden laatuparametrit olivat molemmissa tutkimuksissa kokonaiskloori, sähkönjohtavuus, pH, veden lämpötila, orgaanisen hiilen määrä (TOC) sekä sameus. Tapahtumat simuloitiin sattumanvaraisesti jokaiselle veden laatuparametrille. Tämän jälkeen simuloidut tapahtumat laitettiin normaalin datan päälle kuvan 7 osoittamalla tavalla. On huomattava, että MVE-lajittelumallin tapauksessa simuloituja tapahtumia käytettiin vasta mallin koulutuksen jälkeen, kun taas SVE:n tapauksessa ne olivat koulutuksessa mukana. (Oliker & Ostfeld, 2014a; Oliker & Ostfeld, 20014b)

Li et al. CCA-artikkelissa data oli laboratoriossa suoritetusta pilottikokeesta 150 minuutin ajalta ja mittaukset tehtiin minuutin välein. Ensimmäiset 120 minuuttia mittauksista olivat normaalista datasta. 120 minuutin kohdalla normaaliin veteen lisättiin akryyliamidia ja loput 30 minuuttia mittauksista olivat kontaminaatiosta. Tässä artikkelissa seurattavia veden laatuparametrejä oli 8: veden lämpötila, pH, sameus, sähkönjohtavuus, redox-potentiaali, UV-254, nitraatti-/typpipitoisuus ja fosfaattipitoisuus. (Li et al., 2016) Normaalit mittaustulokset ja kontaminaation jälkeiset mittaustulokset näkyvät kuvassa 8. Kuvasta nähdään, että akryyliamidin

lisäyksen jälkeen erityisesti lämpötila, redox-potentiaali, nitraatti-/typpipitoisuus, UV-254 ja fosfaattipitoisuus reagoivat vieraan aineen läsnäoloon.



Kuva 7. *MVE-lajittelumallia varten simuloidut tapahtumat. Kuvassa normaali ajo näkyy sinisenä ja tapahtumat punaisena. (Muokattu lähteestä Oliker & Ostfeld, 2014a)*



Kuva 8. *Laboratoriossa simuloitu kontaminaatio CCA:n testausta varten. Kuvassa punaisen viivan vasemmalla puolella a-osiossa näkyy normaalin ajon aikainen data. Punaisen viivan kohdalla veteen on lisätty akryyliamidia ja oikealla b-osiossa data on kontaminaation ajalta. (Muokattu lähteestä Li et al., 2016)*

Sekä kuvassa 7, että kuvassa 8 näkyy tapahtumien lisäksi myös poikkeamia perusviivasta, jotka eivät johdu simuloidusta kontaminaatiosta. EDS:n ei tulisi huomioida näitä poikkeamia.

4.2.2 Saadut tulokset

Sekä Olikerin ja Ostfeldin että Li et al. EDS:ien suorituskykyä testattiin useita kertoja. Olikerin ja Ostfeldin kokeissa tapahtumien kestoa ja poikkeavuutta normaalista datasta vaihdeltiin eri testikerroilla. Li et al. tutkimuksessa lisätyn akryyliamidin määrä vaihteli. (Oliker & Ostfeld, 2014a; Oliker & Ostfeld, 20014b; Li et al., 2016) Selkeyden vuoksi tästä eteenpäin puhutaan kontaminaation voimakkuudesta. Mitä voimakkaampi kontaminaatio on, sitä pidempään se kestää ja sitä enemmän mittaustulokset poikkeavat perusviivasta.

Olikerin ja Ostfeldin artikkeleissa SVM:stä ja MVE:stä kontaminaation voimakkuudet on jaettu kolmeen ryhmään. Annetaan ryhmille nimeksi pienimmästä voimakkuudesta suurimpaan ”heikko”, ”keskivoimakas” ja ”voimakas”. Ryhmien tarkemmat jaotteluperusteet löytyvät Olikerin ja Ostfeldin artikkeleista (Oliker & Ostfeld, 2014a; Oliker & Ostfeld, 2014b). CCA:ta käsittelevässä artikkelissa kontaminoituneen veden akryyliamidikonsentraatioita oli neljä: 1,600 mg/l; 3,200 mg/l; 4,800 mg/l ja 6,400 mg/l (Li et al., 2016). Jotta kaikkia kolmea EDS:ää voidaan paremmin vertailla, luokitellaan CCA-artikkelin kontaminaatioiden voimakkuudet seuraavasti: 1,600 mg/l akryyliamidikonsentraatio on heikko, 3,200 mg/l akryyliamidikonsentraatio on keskivoimakas ja 4,800 mg/l sekä 6,400 mg/l akryyliamidikonsentraatiot ovat voimakkaita. Kaksi suurinta akryyliamidikonsentraatiota on vertailun helpottamiseksi merkattu voimakkaaksi. Kuitenkin jatkossa verrattaessa CCA:ta SVM:ään ja MVE:hen on huomioitava, että jo CCA-artikkelissa käytetyn datan takia CCA ei ole täysin vertailukelpoinen muihin käsiteltäviin EDS:iin.

Taulukkoon 1 on kerätty CCA:n, SVM:n ja MVE:n suorituskyvyt (PD ja Oikeellisuus) jokaisella kontaminaation voimakkuudella. Lisäksi taulukosta näkee algoritmien suorituskykyjen testaamiseen käytetyn datan määrä ja monitoroidut veden laatuparametrit. Taulukossa CCA:n arvot on laskettu Li et al. artikkelissa suoritetun kahden kokeen keskiarvona ja lisäksi oikeellisuuden laskemiseen on käytetty kaavaa (3) (Li et al., 2016). SVM:n ja MVE:n tulokset ovat Olikerin ja Ostfeldin artikkelista. Jokaisella voimakkuudella suorituskyvyn arvot ovat SVM:n ja MVE:n tapauksessa kymmenen kokeen keskiarvoja. (Oliker & Ostfeld, 2014c) CCA:n, SVM:n ja MVE:n suorituskykyyn liittyviä tutkimuksia ei tällä hetkellä löydy kuin yksi per EDS.

Taulukossa on myös esitetty Pearsonin korrelaatioon ja euklidiseen etäisyyteen (Pearson correlation Euclidean distance, PE) perustuvan menetelmän suorituskykyjä. PE:n suorituskykyä kuvaa oikeellisuuden sijaan FAR-luku (kaava 2). Taulukon PE:n tulokset ovat Liu et al. artikkeleista (Liu et al., 2014; Liu et al., 2015a; Liu et al. 2015b & Liu et al., 2016).

Taulukko 1 *CCA:n, SVM:n ja MVE:n suorituskyvyn testauksessa käytetyn datan määrä, monitoroidut veden laatuparametrit sekä algoritmien PD-arvot ja oikeellisuus kolmella eri kontaminaation voimakkuudella. Lisäksi PE:n monitoroidut laatuparametrit, PD ja FAR. PD-arvo kuvaa algoritmin kykyä havaita kontaminaatio. Oikeellisuus kuvaa algoritmin kykyä tulkita aika-askelia oikein. PD:n ja oikeellisuuden tulisi olla mahdollisimman lähellä yhtä. FAR kertoo EDS:n todennäköisyyden antaa väärä hälytys ja sen tulisi olla mahdollisimman pieni. (Liu et al., 2014[1]; Oliker & Ostfeld 2014a [2]; Oliker & Ostfeld, 2014b [3]; Oliker & Ostfeld 2014c [4]; Liu et al., 2015a[5]; Liu et al. 2015b[6]; Li et al., 2016[7] & Liu et al., 2016[8])*

EDS	Algoritmin testauksessa käytetyn datan määrä	Monitoroidut veden laatuparametrit	Kontaminaation voimakkuus	PD	Oikeellisuus
CCA [7]	Data 150 minuutin ajalta minuutin välein mitattuna.	Veden lämpötila, pH, sameus, sähkönjohtavuus, redox-potentiaali, UV-254, nitraatti-/typpipitoisuus ja fosfaattipitoisuus	Heikko	0,950	0,990
			Keskivoimakas	0,967	0,993
	150 aika-askelta		Voimakas	0,984	0,998
SVM [2,4]	Data 4 viikon ajalta 5 minuutin välein mitattuna.	Veden lämpötila, pH, sameus, sähkönjohtavuus, kokonaiskloori ja TOC	Heikko	0,44	0,87
			Keskivoimakas	0,97	0,92
	8 064 aika-askelta		Voimakas	0,98	0,94
MVE [3,4]	Data 4 viikon ajalta 5 minuutin välein mitattuna.	Veden lämpötila, pH, sameus, sähkönjohtavuus, kokonaiskloori ja TOC	Heikko	0,61	0,92
			Keskivoimakas	0,99	0,95
	8 064 aika-askelta		Voimakas	1,00	0,96
					FAR
PE[1,5,6,8]		Veden lämpötila, pH, sameus, sähkönjohtavuus redox-potentiaali UV-254, nitraatti-/typpipitoisuus ja fosfaattipitoisuus		0,76 – 1,00	0,00 – 0,87

Dejus et al. artikkelissa on taulukon 1 EDS:ien suorituskykyjen lisäksi koottu useamman EDS:n suorituskykyjä. Tässä työssä tarkasteltuihin EDS:iin verrattuna suorituskyvyn puolesta parhaiten pärjäsi Pearsonin korrelaatioon ja euklidiseen etäisyyteen perustuva menetelmä. Artikkelista nähdään myös, että esimerkiksi EDS:t, jotka hyödyntävät neuroverkkoja, eivät yllä yhtä hyviin suorituskyvyn arvoihin tässä työssä käsiteltävien EDS:ien kanssa. (Dejus et al., 2017)

4.3 Pohdintaa tuloksista

Taulukkoon 1 listatut tulokset viittaavat siihen, että tarkastelluista kolmesta EDS:tä kanonista korrelaatioanalyysiä soveltava EDS on melkein jokaisella osa-alueella paras. Lisäksi vaikka

oikeellisuudesta sitä ei suoraan näekään, CCA ei aiheuttanut testeissä ainuttakaan väärää hälytystä (Li et al., 2016). Ainoastaan MVE ylsi parempiin PD-arvon tuloksiin, kun kyseessä oli keskivoimakas ja voimakas kontaminaatio. Voimakkuuden laskiessa heikkoon, MVE:n ja SVM:n kyky havaita tapahtuma laski huomattavasti.

CCA:ta on kuitenkin verrattava muihin EDS:iin varautuen. Ensinnäkin Li et al. CCA:ta käsittelevässä artikkelissa tutkittiin tapahtuman havaitsemista raakavedestä, kun taas SVM ja MVE monitoroivat verkostossa olevaa talousvettä. Seurattavat veden laatuparametrit ovat kaikilla EDS:illä melkein samat, mutta erojakin niistä löytyy. Raakavedessä ei pitäisi olla vapaata klooria ja siksi sitä ei CCA:n tapauksessa ole mitattu. Lisäksi CCA:n testauksessa käytetty raakavesi on MVE:n ja SVM:n käyttämää vettä sameampaa ja siinä saattaa esiintyä suurempia lämpötilan vaihteluita vuorokauden ajan mukaan. (Oliker & Ostfeld, 2014a; Oliker & Ostfeld, 2014b; Li et al., 2016) Mikäli CCA otettaisiin käyttöön talousveden monitorointiin, tarvitsisi miettiä uudestaan mitä veden laatuparametrejä käytetään ja miten algoritmi tulisi optimoida erilaiselle mittausdatalle. Oletettavasti SVM ja MVE olisi helpompi muuttaa monitoroimaan raakavettä, sillä koneoppimisalgoritmit suorittaisivat harjoitusvaiheessa itse optimoinnin uuteen ympäristöön.

Toinen huomioitava asia verrattaessa CCA:ta SVM:ään ja MVE:hen on suorituskyvyn testaamiseen käytetyt koejärjestelmät ja datan määrä (taulukko 1). CCA:n data tuli vain 150 minuutin ajalta ja koe toistettiin ainoastaan kaksi kertaa (Li et al., 2016). Muut tarkasteltavat EDS:t käyttivät dataa kuukauden ajalta ja testit toistettiin kymmenen kertaa jokaiselle kontaminaation vahvuudelle (Oliker & Ostfeld, 2014a; Oliker & Ostfeld, 2014b). Lisäksi koko CCA:n data oli laboratoriokokeesta, kun taas Oliker ja Ostfeld käyttivät normaalina datana oikeaa dataa talousvesiverkosta testatessaan SVM:ää ja MVE:tä. Kun ottaa vielä huomioon sen, että SVM:n ja MVE:n luomia malleja voi vielä käyttöönoton jälkeen päivittää uudella datalla, CCA ei välttämättä ole joka tilanteessa paras vaihtoehto. (Oliker & Ostfeld, 2014a; Oliker & Ostfeld, 2014b; Li et al., 2016)

Myös suorituskyvyn testaamiseen käytetyissä simuloiduissa kontaminaatioissa on eri EDS:ien kohdalla eroja. CCA:ssa käytetty oikea akryyliamidin injektio veteen todistaa, että algoritmi kykenee tunnistamaan ainakin kyseisen aineen tapauksessa oikean kontaminaation (Li et al., 2016). Tarvitaan kuitenkin vielä paljon tutkimusta CCA:n suorituskyvystä erilaisten vaarallisten aineiden kanssa. Tämä voi olla haastavaa ja kallista, sillä ihmisen terveydelle haitallisia aineita on paljon, eikä niistä kaikkia välttämättä osata ottaa jatkotutkimuksissa huomioon.

SVM:n ja MVE:n kykyä havaita kontaminaatio ei testattu millään tietyllä aineella (Oliker & Ostfeld, 2014a; Oliker & Ostfeld, 2014b). Haittapuolena tässä on se, että luodut kontaminaatiot eivät välttämättä vastaa todellisia kontaminaatioita (Oliker & Ostfeld, 2014c). Tilanteen voi toisaalta nähdä myös positiivisena: Koska jokaisen veden laatuparametrin reagointia jokaisen vaarallisen aineen läsnäoloon ei tunneta, on hyvä, että kyseiset algoritmit pystyvät tunnistamaan myös täysin sattumanvaraisesti luotuja kontaminaatioita. Tällöin EDS tunnistaa mahdollisesti myös täysin sattumanvaraiset vieraat aineet vedestä. Myös jatkotutkimus on oletettavasti SVM:n ja MVE:n tapauksessa halvempaa, sillä uusia keinotekoisia kontaminaatioita pystytään luomaan helposti.

Yhteenvetona CCA:sta, SVM:stä ja MVE:stä puhtaasti ilmoitettuja suorituskykyjä tarkastelemalla (taulukko 1) paras EDS on CCA. Menetelmä vaatii kuitenkin paljon jatkotutkimusta, jotta tuloksista saadaan luotettavampia ja menetelmä voidaan ottaa käyttöön nimenomaan talousveden monitoroinnissa. SVM:stä ja MVE:stä paremmin suoriutui MVE. Paremman suorituskyvyn lisäksi MVE on yksinkertaisempi, muun muassa siksi, että sen harjoitus tapahtuu ohjaamattomasti (Oliker & Ostfeld, 2014c). Mallin opetukseen riittäisi siis periaatteessa vain data talousvesiverkoston normaalista ajosta. Voi siis olla, että SVM:ää ei kannata enää jatkojalostaa talousvesiverkoston EDS:ksi.

4.4 Pohdintaa EDS:ien soveltuvuudesta talousvesiverkoston valvontaan

Luvussa 4.2 esitetyt tulokset antavat sellaisen kuvan, että tarkastellut EDS:t ovat toimivia ja ne voisivat lisätä huomattavasti talousvesijärjestelmän turvallisuutta. Vaikka menetelmät vaativat lisätutkimusta, niistä voi olla tulevaisuudessa hyötyä, kunhan EDS:n käyttöönottokustannukset ja huolto ei ole liian kallista. Tarkasteltujen EDS:ien kykyä havaita kontaminaatio talousvedestä voi kuvitella olettamalla, että talousvesilaitos toimii viidenkymmenen vuoden ajan ja sinä aikana vesi kontaminoituu kerran. Taulukossa 2 esitettyjen PD-arvojen keskiarvo on 0,88, jolloin tässä tapauksessa on 88% todennäköisyys, että vesilaitoksen toiminta-ajan ainoa kontaminaatio havaitaan.

Kuitenkaan hyvä kontaminaation havaitsemistodennäköisyys ei yksin riitä. On tärkeää, että EDS:n todennäköisyys antaa väärä hälytys on erittäin pieni. Toisin kun kontaminaatioita, EDS joutuu käsittelemään normaaleja mittaustuloksia muutaman minuutin välein. Oletetaan nyt, että EDS:n todennäköisyys antaa normaalin mittaustuloksen kohdalla väärä hälytys on 0,001. Jos mittauksia tehdään viiden minuutin välein, EDS antaa silti väärän hälytyksen keskimäärin noin 3,5 päivän välein. Vesilaitoksen tai muun päättävän tahon on siis itse päätettävä kuinka suuri FAR-arvo ja kuinka pieni PD-arvo sallitaan.

5. YHTEENVETO JA JOHTOPÄÄTÖKSET

Työssä tutkittiin kolmen erilaisen EDS:nä käytetyn algoritmin soveltuvuutta monitoroimaan talousvettä kontaminaatioiden varalta. Tarkasteltavat algoritmit olivat painotettu tukivektorikone (SVM), pienimmän tilavuuden ellipsoidi -lajittelumalli (MVE-lajittelumalli) ja kanoninen korrelaatioanalyysi (CCA). Näistä SVM ja MVE ovat koneoppimisalgoritmeja ja CCA on tilastollinen tapa laskea korrelaatio uuden mittauksen ja vanhojen mittausten välillä. Algoritmien suorituskyvyn testaamisessa saadut tulokset taulukoitiin ja tuloksia verrattiin toisiinsa. Eri EDS-sovellutusten vertailussa huomioitiin myös eroavaisuudet datassa, jota käytettiin EDS:n testaamiseen ja kouluttamiseen.

Ilmoitettuja suorituskykyjen tuloksia tarkastelemalla SVM:stä, MVE:stä ja CCA:sta paras EDS on CCA. CCA ei koko testauksen aikana aiheuttanut yhtäkään väärää hälytystä ja heikollakin kontaminaation voimakkuudella CCA havaitsi kontaminaation 95% todennäköisyydellä. Parhaimmillaan CCA ylsi 98,4%:n tapahtuman havaitsemistodennäköisyyteen. Kuitenkin toisin kuin SVM:n ja MVE:n testaamisessa, CCA:n käyttämä data ei ollut peräisin oikeasta talousvesiverkostosta. CCA:n testaukseen käytetty data tuli laboratoriossa suoritetusta pilottikokeesta ja käytetyn datan määrä oli huomattavasti pienempi kuin SVM:n ja MVE:n testauksessa. CCA:n testaukseen käytettiin vain 150 mittausta, kun taas SVM:n ja MVE:n testaukseen käytettiin 8 064 mittausta. Lisäksi CCA:n data mukaili raakavettä talousveden sijaan.

Ottaen huomioon testauksessa käytetyn datan määrän ja mahdollisuuden päivittää mallia käytön aikana, paras EDS saattaisi olla MVE. MVE aiheutti enemmän vääriä hälytyksiä kuin CCA, mutta parhaimmillaan havaitsi kontaminaatioita paremmalla todennäköisyydellä. Keskivoimakkaalla kontaminaatiolla MVE:n PD oli 0,99 ja voimakkaalla kontaminaatiolla MVE havaitsi jokaisen tapahtuman. Heikolla kontaminaation voimakkuudella MVE ei kuitenkaan havainnut tapahtumia yhtä hyvin ja sen PD oli vain 0,61. SVM oli jokaisella osa-alueella huonompi kuin MVE, joten tulevaisuudessa on tuskin järkevää panostaa SVM:n jatkokehitykseen.

Työn yhtenä tavoitteena oli selvittää, sopivatko tarkasteltavat EDS:t sellaisenaan käytettäväksi talousveden monitorointiin. Työssä esitettyjen suorituskykyjen perusteella, mikäli tarkasteltuja EDS:iä aiotaan ottaa käyttöön, ne tarvitsevat vielä lisää tutkimusta ja testausta. EDS:ien PD-arvot ja oikeellisuudet ovat vielä oletettavasti liian pieniä, jotta vesilaitosten kannattaisi investoida niihin. CCA ja MVE vaikuttavat kuitenkin lupaavilta ratkaisuilta talousvesiverkoston monitorointiin. Erityisesti CCA vaikuttaa lupaavalta, jos testauksissa osoittautuu, että CCA:n suorituskyky pysyy samankaltaisena useammilla talousvedelle vierailla aineilla testattuna. CCA vaatii myös lisää tutkimusta, mikäli se aiotaan ottaa käyttöön talousveden monitorointiin raakaveden monitoroinnin sijaan. Jos MVE:n suorituskyky paranee vielä jatkotutkimuksissa, sekin on varteenotettava vaihtoehto. Ohjaamattomasti oppivana koneoppimisalgoritmina MVE:n koulutus on yksinkertaista ja sen kouluttaminen on mahdollista ilman dataa kontaminaatiosta.

Tulevaisuudessa, valittaessa sopivaa EDS:ää, vesilaitosten tai muiden vastaavien tahojen on pohdittava, onko heille tärkeämpää korkea todennäköisyys kontaminaation havaitsemiselle vai

vähäinen väärien hälytysten määrä. Väärät hälytykset voivat aiheuttaa kustannuksia, mutta jos vesilaitos voi esimerkiksi jotenkin ehkäistä näitä kustannuksia, voi olla parempi valita EDS, joka havaitsee vedestä vieraat aineet paremmin. Vesilaitoksen on pohdittava myös muun muassa EDS:n testauksen ja kouluttamisen monimutkaisuutta. Ohjaamattomasti oppiva koneoppimisalgoritmi voi olla harjoitusvaiheen yksinkertaisuuden takia houkuttelevampi vaihtoehto kuin suorituskyvyltään parempi, mutta monimutkainen EDS.

LÄHTEET

Alpaydin, E. (2014). Introduction to Machine Learning. The MIT Press, Third edition, pp. 1–13.

Banko, M. & Brill, E. (2001). Scaling to Very Very Large Corpora for Natural Language Disambiguation. Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, pp. 26–33.

Boser, B.E., Guyon, I.M. & Vapnik, V.N. (1992). A Training Algorithm for Optimal Margin Classifiers. Proceedings of the Fifth Annual Workshop on Computational Learning Theory. ACM, pp. 144–152.

Byer, D. & Carlson, K.H. (2005). Real-time Detection of Intentional Chemical Contamination in the Distribution System. Journal American Water Works Association, Vol. 97(7), pp. 130–133.

Cortes, C. & Vapnik, V.N. (1995). Support Vector Networks. Machine Learning, Vol. 20(3), pp. 273–297.

Dejus, S., Neščerecka, A. & Juhna, T. (2017). On-line Drinking Water Contamination Event Detection Methods. Environment. Technology. Resources. Proceedings of the 11th International Scientific and Practical Conference, Vol. 1, pp. 77–81

Géron, A. (2017). Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, First edition. O'Reilly Media, Inc, pp. 3–31. Saatavissa: http://files.isec.pt/DOCUMENTOS/SERVICOS/BIBLIO/INFORMA%C3%87%C3%95ES%20ADICIONAIS/Hands-on-machine_Gueron.pdf. Haettu 31.5.2019.

Hall, J., Herrmann, J.G., Kefauver, P.C., Krishnan, E.R., Marx, R.B. & Zaffiro, A.D. (2007). Online Water Quality Parameters as Indicators of Distribution System Contamination. Journal American Water Works Association, Vol. 99(1), pp. 66–77

Hasan, J. (2005). Technologies and Techniques for Early Warning Systems to Monitor and Evaluate Drinking Water Quality: A State-of-the-Art Review. U.S. EPA., Office of Water, pp. 9–18.

Härdle, W. & Simar, L. (2007). Applied Multivariate Statistical Analysis. Springer. pp. 321–330

Li, R., Liu, S., Smith, K. & Che, H. (2016). A Canonical Correlation Analysis Based Method for Contamination Event Detection in Water Sources. Environmental Science: Processes and Impacts, Vol. 18(6), pp. 658–666.

Liu, S., Che, H., Smith, K. & Chen L. (2014). Contamination Event Detection Using Multiple Types of Conventional Water Quality Sensors in Source Water. Environmental Science: Processes & Impacts, Vol. 16(8), pp. 2028–2038.

Liu, S., Che, H., Smith, K., Lei, M. & Li, R. (2015a). Performance Evaluation for Three Pollution Detection Methods Using Data from a Real Contamination Accident. Journal of Environmental Management, Vol. 161, pp. 385–391.

Liu, S., Smith, K. & Che, H. (2015b). A Multivariate Based Event Detection Method and Performance Comparison with Two Baseline Methods. Water Research, Vol. 80, pp. 109–118.

Liu, S., Li, R., Smith, K. & Che, H. (2016) Why Conventional Detection Methods Fail in Identifying the Existence of Contamination Events. Water Research, Vol. 93, pp. 222–229.

Louridas, P. & Ebert, C. (2016). Machine Learning. IEEE Software, Vol. 33(5), pp. 110–115.

Mounce, S.R., Mounce, R.B., Jackson, T., Austin, J. & Boxall, J.B. (2014). Pattern Matching and Associative Artificial Neural Networks for Water Distribution System Time Series Data Analysis. Journal of Hydroinformatics, Vol. 16(3), pp. 617–632.

Murray, S., Ghazali, M. & McBean, E.A. (2011). Real-Time Water Quality Monitoring: Assessment of Multisensor Data Using Bayesian Belief Networks. Journal of Water Resources Planning and Management, Vol. 138(1), pp. 63–70.

Oliker, N. & Ostfeld, A. (2014a). A Coupled Classification – Evolutionary Optimization Model for Contamination Event Detection in Water Distribution Systems. Water Research, Vol. 51, pp. 234–254.

Oliker, N. & Ostfeld, A. (2014b). Minimum Volume Ellipsoid Classification Model for Contamination Event Detection in Water Distribution Systems. Environmental Modelling & Software, Vol. 57, pp. 1–12.

Oliker, N. & Ostfeld, A. (2014c). Comparison of Multivariate Classification Methods for Contamination Event Detection in Water Distribution Systems. Procedia Engineering, Vol. 70, pp. 1271–1279.

Onnettomuustutkintakeskus. (2007). Puhdistetun jäteveden joutuminen talousvesiverkostoon Nokialla 28.–30.11.2007. Tutkintaselostus, B2/2007Y.

Rousseeuw, P.J., (1985). Multivariate Estimation with High Breakdown Point. Mathematical Statistics and Applications, Vol. B, pp. 283–297.

Spence, S., Rosen, J.S. & Bartrand T. (2013). Using Online Water Quality Data to Detect Events in a Distribution System. Journal - American Water Works Association, Vol. 105(7), pp. 22–26.

U.S. EPA. (2010). Water Quality Event Detection Systems for Drinking Water Contamination Warning Systems: Development, Testing, and Application of CANARY. Environmental Protection Agency, pp. 2–4.

Whalen, P., McBean, E.A., Journal, K. & Ghazali, M. (2010). Supporting a Drinking Water Contaminant Warning System Using the Adenosine Triphosphate Test. Canadian Journal of Civil Engineering, Vol. 37(11), pp. 1423–1431.

Zhao, H., Hou, D., Huang, P. & Zhang, G. (2014). Water Quality Event Detection in Drinking Water Network. Water, Air & Soil Pollutin, Vol. 225(11), pp. 1–15.