Kirjoitelma

Tietojenkäsittelytieteen kandiohjelma

Koneoppimisen menetelmät ja käyttökohteet lääketutkimuksessa -ja kehityksessä

Heikki Pulli

26.9.2021

Matemaattis-luonnontieteellinen tiedekunta

Helsingin yliopisto

Ohjaaja(t)

Tarkastaja(t)

Yhteystiedot

PL 68 (Pietari Kalmin katu 5)
00014 Helsingin yliopisto

Sähkopostiosoite: info@cs.helsinki.fi
URL: http://www.cs.helsinki.fi/

Sisällys

1 Johdanto

Koneoppimismallien käyttö lääketutkimuksessa ja -kehityksessä on lisääntynyt. Lisäksi tutkimus, jossa selvitetään, kuinka eri koneoppimismalleja voidaan hyödyntää lääkekehityksen tarpeisiin on lisääntynyt. Esimerkiksi jotkin organisaatiot järjestävät kilpailuja ja tapahtumia, joiden tarkoituksena on yrittää lötää uusia käyttökohteita koneoppimismalleille lääketutkimuksen tutkimusalueelta. Yrityksille kannustimena käyttää koneoppimismalleja lääkkeiden kehityksessä toimii mahdollisuus vähentää lääkkeiden kehittämiseen käytettyä aikaa ja resursseja, kun koneet voivat antaa omat ennusteensa, millä tutkittavilla lääkkeillä on suurimmat todennäköisyydet päästä uusilta lääkkeiltä vaadituista testeistä läpi ja tuotantoon. Lisäksi jotkut edistyneet mallit voivat annetusta datasta päätellä, mistä yhdisteistä muodostuu tiettyyn sairauteen tehoava lääke ja mitkä ovat tämän lääkkeen valmistusvaiheet. Nämä ohjeet kone voi sitten antaa lääkkeenkehittäjille, jotka voivat validoida koneen antaman yhdisteen toimivuuden ja vaiheiden pätevyyden. Nämä ovat vain muutamia esimerkkejä, kuinka koneoppimista voidaan hyödyntää, mutta käyttökohteita on useampia. Koneoppimisen hyödyntämisesssä lääketutkimuksessa on kuitenkin vielä monia puutteita ja heikkouksia. Suurin osa puutteista liittyy datan saatavuuteen ja saatavilla olevan datan käytettävyyteen. Kaikki saatavilla oleva data ei ole aina tilannekohtaisesti tarpeeksi hyvää, jotta sitä voitaisiin käyttää haluttuun tilanteeseen. Lisäksi hyvän datan tuottamista voi myös paikoin hidastaa sen korkea tuotantokustannus. Lisäksi eri mallien tulosten validointia vaikeuttaa se, että ei voida koskaan täsyin tietää, miten kone on tulokseensa päätynyt. Nämä ovat ongelmia, joihin tieteellinen yhteisö yritää saada ratkaisuja [1]. Näistä puutteista huolimatta koneoppiminen on todistanut olevansa tehokas työkalu lääkekehityksessä.

2 Lääketutkimuksessa -ja kehityksessä käytetyt koneoppimisen mallit

Lääketutkimuksessa käytetään monia eri koneoppiminen malleja. Eri mallit kuitenkin soveltuvat paremmin eri tilanteisiin, jolloin tutkija -tai kehittäjäryhmän tulee valita käytettävä malli ongelman mukaisesti. On kuitenkin todettu, että eri paradigmaa noudattavat mallit sopeutuvat tiettyihin tehtäviin paremmin.

On huomattu, että ohjatulla oppimisella ja vahvitusoppimisella koulutetut mallit soveltuvat paremmin tehtäviin, joissa on tarkoituksena luokitella dataa tutkittavien ominaisuuksien perusteella. Näitä ovat esimerkiksi kuviin perustuva diagnoosi tai syöpään vaikuttavien geenien RNAi seulonnassa. Käytettyjä ohjattuja koneoppimismalleja ovat esimerkiksi lineaari regressio ja syvät neuroverkot.

Ohjaamatonta koneoppimisen malleja käytetään uusien ennaltatuntemattomien asioiden löytämiseen saatavilla olevasta datasta. Tällaisia asioita voivat olla esimerkiksi uusien biomarkkereiden etsintä. Tällöin datasta tavallisesti etsitään ryppäitä, joita voidaan analysoida. Ohjaamattomia koneoppimismalleja ovat puolestaan k-lähin klusterointi ja itseohjautuva kartta tai Kohosen kartta.

Kaikki käytettävät mallit ovat kuitenkin vain niin hyviä kuin saatavilla oleva data. Saatavilla olevan datan tulee olla tarkkaa, tarpeeksi kuvaavaa ja vertaisarvioitua. Kuitenkin tarvittavan datan määrä riippuu käytettävästä mallista ja tutkittavasta asiasta. Tutkittavasta asiasta riippuen tutkimukseen saattaa liittyä myös käytettävän datan tuottaminen. Paras data on systemaattisesti tuotettua jossa on ollut mukana mahdollisimman vähän muuttujia ja joka on kuvaavasti nimetty ja luokiteltu.

3 Koneoppimismallien käyttökohteet

Lääketutkimuksen tavoitteena on kehittää lääkkeitä, jotka vaikuttavat tautiin muokkaamalla haluttua molekyylitason kohdetta. Koska saatavilla olevaa dataa on enenemissä määrin, joka kuvaa tauteja ja mihin ne vaikuttavat ja mikä niihin vaikuttaa, niin koneoppimisen menetelmiä voidaan hyödyntää mahdollisten uusien tautien ja niihin tehoavien lääkkeiden tutkimuksessa.

Yksi koeneoppimisen käyttökohteita on ollutkin taudin aiheuttajan tunnistaminen. Tätä voidaan tutkia analysoimalla asioita, joihin tauti vaikuttaa. Kun on olemasssa dataa siitä, mitkä mahdolliset asiat vaikuttavat niihin asioihin, joihin tauti vaikuttaa, niin pystytään rajaamaan taudinaiheuttajia.

Koneoppimisen menetelmiä käytetään myös saatavilla olevan lääketieteellisen kirjallisuuden analysointiin. NLP -menetelmien avulla voidaan seuloa suuresta määrästä tieteellisiä artikkeleita vain ne, jotka ovat tarpeellisia tilanteeseen. Tämä kuitenkin on riippuvaista siitä, kuinka hyvin julkaisut on annotoitu aiheidensa mukaan.

Yksi keskeisimmistä koenoppimisen käyttökohteita on ennustaa, kuinka suurella todennäköisyydellä kehitettävä lääke pääsee kliinisiin testeihin ja niistä läpi. Roullardin tutkimusryhmä tutki lääkkeiden onnistumista käyttämällä koneoppimisen menetelmiä. Tutkimuksissa käytettiin dataa, joka kertoo, onko lääke päässyt testeistä läpi vai ei ja mikä on ollut lääkkeen vaikuttava tekijä. Roullardin tutkimusryhmä päätyi lopputulokseen, että geeniekspressiivinen datapystyi parhaiten ennustamaan lääkkeen onnistumisen.

Koneoppimista voidaan hyödyntää myös pienten molekyylien vaikutusten ennustamisessa. On huomattu, että multi-task DNN ovat tähän tarkoitukseen tehokkaampia, kuin aikaisemmin käytetyt menetelmät. Käytetty One-shot tekniikka tarvitsee vähämmen dataa ja aikaa ennustaakseen tietyn yhdisteen reaktion tietyissä olosuhteisssa. Multi-task mallit ovat kuitenkin Single-task malleihin verrataen paljon enemmän riippuvaisempia datasta.

Neuroverkkoja ja nykyisiä puuhaku algoritmeja voidaan myös hyödyntää lääkkeen syntetisoinnin välivaiheiden suunnittelussa. Tämä voidaan toteuttaa käyttämällä dataa synteettisestä kemiasta. Tämä on kuitenkin vaikea prosessi, koska tiedon määrä eri reaktioista kasvaa eksponentiaaliesti ja ei aina ole tietoa kaikista tilanteeseen liittyvisä reaktiosta. Seglerin tutkimusryhmä käytti Monte carlo -puuhakua ja neuroverkkoa apuna ohjatak-

seen haku algoritmia oikeaan suuntaan. Tutkimusryhmä käytti dataa Reaxys tietokannasta, joka sisältää tietoa eri yhdisteistä ja niiden välisistä reaktioista. Tämä menetelmä oli kolmekymmentä kertaa nopeampi kuin aikaisemmat menetelmät ja tulokset olivat keskivertaisesti yhtä hyvät kuin tutkijan itse tekemä menetelmä luoda yhdiste synteettisesti.

Konenoppimista käytetään myös kuva-analyysissa. Tätät käytetään varsinkin patologisessa kontekstissa, kun halutaan tutkia esimerkiksi miten jokin tauti ilmenee näytteessä tai miten jokin lääke on vaikuttanut näytteeseen.

4 Koneoppimismallien heikkoudet ja puutteet

Vaikka koneoppiminen vaikuttaa suoraviivaistavan ja nopeuttavan lääkkeiden tutkimusta ja -kehitystä, niin malleissa on kuitenkin myös heikkouksia ja puutteita. Yksi näistä on mallien tulosten tulkittavuus ja varmennettavuus. Koneista nähdään vain, mitä dataa mallille annetaan ja minkä tuloksen se antaa käyttäjälle. Miten kone päätyi lopputulkseensa on liki mahdoton saada selville. Tämä on yksi hidastava aspekti suuremmassa koneoppimisen mallien käyttöönotossa.

Toinen ongelma on koneiden tulosten toistettavuus. Koneiden antamat tulokset ovat hyvin riippuvaisia koneen alkuasetuksista ja osittain jopa siitä, missä järjestykssä data koneelle annetaan, vaikka kone saisi täysin saman datan joka tilanteessa.

Hyvän datan puute on myös ongelma. Vaikka dataa olisikin paljon johonkin tilanteeseen, niin datan heikko laatu saattaa kuitenkin aiheuttaa ongelmia. Koska laadukas data on systemaatiisesti tuotettua, hyvin annotoitua ja vertaisarvioitua, niin datan luomisessa saattaa kestää hyvinkin kauan aikaa ja datan tuotantokustannus saattaa nousta hyvinkin suureksi.

Ongelmana on myös tietävän ja osaavan henkilöstön puute. Koska tutkimuksissa käsitellään sekä tietojenkäsittelytieteiden aiheita että biologian ja lääketutkimuksen aiheita, niin osaavaa henkilöstöä on hankalampi löytää.

Lähteet

[1] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer ja S. Zhao. ”Applications of machine learning in drug discovery and development”. eng. *Nature reviews. Drug discovery* 18.6 (2019), s. 463–477. ISSN: 1474-1776.