DUOMENŲ TYRYBA MEDICINOJE: TAIKYMAS, PROBLEMOS IR GALIMYBĖS

Olegas Niakšu
Vilniaus universiteto Matematikos ir informatikos institutas

Santrauka

Pagrindimas. Duomenų tyryba, kaip žinių gavimo proceso dalis, išplečia statistinės analizės ribas ir suteikia naujas galimybes tyrėjams ir praktikams analizuojant elektroninėje terpėje saugomus duomenis.

Tikslai ir uždaviniai. Pagrindinis straipsnio tikslas – apžvelgti ir apibendrinti, kaip duomenų tyryba taikoma medicinoje ir sveikatos apsaugos sektoriuje, akcentuojant praktinio taikymo problemines sritis.

Metodai. Atlikta sisteminė literatūros apžvalga, panaudoti kokybinės lyginamosios analizės metodai.

Rezultatai ir apibendrinimas. Duomenų tyryba sveikatos priežiūros sektoriuje susiduria su specifiniais iššūkiais. Svarbiausi jų: heterogeniški duomenys, pacientų duomenų privatumas, didelės elektroninių duomenų apimtys. Nepaisant to, standartinės duomenų tyrybos taikymo metodikos gali būti naudojamos sveikatos priežiūros srityje. Šiam tikslui pasiūlyta metodika CRISP-DM.

Reikšminiai žodžiai: duomenų tyryba, duomenų gavyba, sveikatos informacinė sistema, medicininė informacinė sistema.

ĮVADAS

Lietuvos sveikatos apsaugos sektoriuje aktyviai taikomos informacinės technologijos. Šiuo metu kuriama nacionalinė elektroninė sveikatos paslaugų ir bendradarbiavimo infrastruktūros informacinė sistema (ESPBI IS), nacionalinė e. recepto sistema bei nacionalinė medicininių vaizdų archyvavimo ir mainų posistemė. Sveikatos priežiūros įstaigos diegia ir tobulina ligoninės informacines sistemas (HIS[1]), radiologinių vaizdų peržiūros ir archyvavimo sistemas (PACS[2] ir RIS), laboratorines informacines sistemas (LIS). Sveikatos informacinėse sistemose (toliau – SIS) kaupiama sustruktūrinta paciento ligos istorija, į kurią įeina klasifikuojami požymiai, tokie kaip diagnozė, demografiniai paciento duomenys, gyvybinės funkcijos, tyrimų rezultatai ir kt. Šių duomenų analizė ir tyryba turi strateginę reikšmę sveikatos sektoriui ir yra svarbi kiekvienam pacientui. Sukauptų duomenų intelektuali analizė siūlo naujas priemones tokiems uždaviniams spręsti: greitesnė pacientų diagnostika, optimalaus gydymo parinkimas, gydymo trukmės ir rezultatų prognozavimas, komplikacijų rizikos nustatymas, sveikatos priežiūros įstaigos išteklių optimizavimas.

Duomenų tyryba yra santykinai nauja sritis. Pastarąjį dešimtmetį aktyviai nagrinėjama, kaip duomenų tyryba taikoma biomedicinoje [1–4]. Mokslinių publikacijų ir pranešimų konferencijose skaičiaus augimas rodo šios tematikos aktualumą. Nors pasaulyje duomenų tyrybos metodai taikomi medicinoje jau ne pirmą dešimtmetį, Lietuvoje tai dar naujovė.

Straipsnio tikslas – apžvelgti duomenų tyrybos sampratą ir pasaulinę patirtį, taikant duomenų tyrybos metodus medicinoje ir sveikatos apsaugos sektoriuje. Taip pat pateikiame metodines gaires, kaip naudoti duomenų tyrybą žinių gavimo procese.

TYRIMO MEDŽIAGA IR METODAI

Atlikta literatūros šaltinių, nagrinėjančių duomenų tyrybos taikymus medicinoje, apžvalga. Pateikiama agreguota ir struktūrinta informacija apie duomenų tyrybos taikymo uždavinius, taikomus metodus bei jų naudojimo metodiką.

Literatūros šaltiniai atrinkti pagal šiuos paieškos kriterijus: „duomenų tyryba medicinoje", „duomenų tyryba sveikatos sektoriuje", „biomedicininių

[1] HIS, angl. *hospital information system*, ligoninės informacinė sistema.

[2] PACS, angl. *Picture archyve system*, RIS, angl. *radiology information system.*

Adresas susirašinėti: Olegas Niakšu
Vilniaus universiteto
Matematikos ir informatikos institutas
Akademijos g. 4, 08663 Vilnius
El. p. niaksu@acm.org

duomenų klasifikacija", „biomediciniinių duomenų analizė", „medicinos statistika", „medicinos informacinių sistemų duomenų analizė", „medicininės ontologijos" ir derinant išvardytus terminus. Didelė informacijos dalis atrinkta *ScienceDirect* ir *MedLink* duomenų bazėse. Prioritetas buvo teikiamas šaltiniams, publikuotiems nuo 2005 metų. Taip pat panaudotos senesnės publikacijos, kuriose apžvelgiami fundamentalūs duomenų analizės ir tyrybos aspektai. Anot M. Stacey ir kt. [5], daugybė mokslinių darbų, kurie sudaro pamatus moksliniam tiriamajam darbui mediciniinių duomenų analizės srityje, buvo publikuoti devyniasdešimtaisiais metais.

Informacija apie elektroniniu būdu kaupiamus duomenis Lietuvos sveikatos priežiūros įstaigose (toliau – SPĮ) gauta bendradarbiaujant su Vilniaus universiteto Santariškių klinikų, LSMU Kauno klinikų bei Klaipėdos universitetinės ligoninės Medicinos statistikos ir IT skyrių darbuotojais.

Duomenų tyryba

Duomenų tyrybos (angl. *data mining*, toliau – DT) terminas nusistovėjo 1990-aisiais, kaip nauja duomenų analizės ir žinių gavimo technologija. Pirma ACM[3] konferencija (SIGKDD) vyko JAV 1995 m., o *Medline* bibliotekos *Medical Subject Headings* (MeSH) terminų žodyne „data mining" buvo užregistruotas 2009 metais.

Žinoma keletas duomenų tyrybos apibrėžimų. Vienas labiausiai paplitusių apibrėžia duomenų tyrybą kaip duomenų analizę, siekiant atrasti nežinomus dėsningumus ir aprašyti (apibendrinti) duomenis suteikiant naujų žinių. Iš šio apibrėžimo išplaukia, kad duomenų tyrybos tikslas – iš duomenų rinkinio (dažnai didelės apimties) gauti naujų žinių ir gilesnį suvokimą, kurie toliau gali būti panaudoti sprendimams priimti.

Gali kilti klausimas, kuo DT skiriasi nuo statistikos? Praktiškai statistiniai metodai dažniausiai taikomi pirminei duomenų analizei, o duomenų tyryba – antrinei duomenų analizei. Yra ir kitų svarbių skirtumų:

- statistikoje suformuluota hipotezė testuojama, pasitelkiant statistinius metodus. DT leidžia taikyti indukcijos metodus, formuluojant hipotezes iš turimų duomenų;
- statistikoje dažniausiai tiriama populiacijos imtis. Duomenų tyryboje, priešingai, dažnai analizuojami visos populiacijos duomenys;
- statistikoje naudojami formalūs matematiniai metodai ir vengiama naudoti netikslius euristinius metodus. DT metodai grindžiami matematika, tačiau joje plačiai taikomos euristikos, lokalaus sprendimo paieškos ir kiti apytiksliai metodai, kurie orientuoti į uždavinius su didele duomenų apimtimi, kategoriniais kintamaisiais arba prasta tiriamų duomenų kokybe.

Medicininė informacija gali būti išreikšta kaip statinė informacija, fiksuojanti momentinę paciento būseną, pavyzdžiui, tyrimų rezultatai, diagnozė; dinaminė – elektrokardiogramos duomenys, grafinė – rentgenogramos, trimatė grafinė – kompiuterinių tomografijų 3D modeliai. Šių įvairialypių duomenų tyryba reikalauja duomenų sąsajumo (angl. *interoperability*) užtikrinimo įrankių, duomenų sandėliavimo (angl. *data warehouse*), vaizdavimo metodų ir specializuotų instrumentų pritaikymo.

Duomenų tyrybos industriniai standartai

Šiuo metu žinoma keletas standartų duomenų tyrybos srityje. Didžiausio dėmesio ir pripažinimo mokslinėse publikacijose bei elektroninėje erdvėje sulaukė CRISP-DM, SEMMA ir PMML standartai. Tai nėra specifiški medicinos duomenų tyrybai skirti standartai, jie taikomi ir kitose srityse. Pirmi du standartai apibrėžia duomenų tyrybos procesą. Platesnė informacija apie atvirą CRISP-DM standartą pateikta skyriuje „Duomenų tyryba kaip žinių gavimo proceso dalis".

PMML[4] yra duomenų tyrybos prognozavimo modeliavimo kalba. PMML yra atviras standartas, leidžiantis formalizuoti prognozavimo modelį ir teikiantis prielaidas DT programinės įrangos interoperabilumui [6]. PMML naudoja XML kalbos sintaksę. Viename PMML dokumente gali būti aprašyti keli DT modeliai.

PMML panaudojimas leidžia tyrėjams viename programinės įrangos pakete sudaryti prognozavimo modelį ir vėliau panaudoti sudarytą modelį kitoje informacinėje sistemoje, pavyzdžiui, ligoninės informacinėje sistemoje arba klinikinių sprendimų palaikymo sistemoje.

Duomenų tyrybos uždaviniai ir metodai

Siekdami sėkmingai panaudoti duomenų tyrybos metodus ir algoritmus, tyrėjai turi gerai suprasti jų taikymo sritis, rūšis ir ypatumus.

[3] ACM, angl. *Association for Computing Machinery*, tarptautinė kompiuterininkų bendruomenė, įsteigta 1947 m. JAV.

[4] PMML, angl. *Predictive Data Mining Markup Language.*

Visas duomenų tyrybos uždavinių sąrašas dar nėra galutinai nusistovėjęs. Dažniausiai šaltiniuose išskiriami šie uždaviniai: klasifikacija, klasterizacija, prognozavimas, asociacija, vizualizavimas, nuokrypių identifikavimas ir ryšių analizė.

DT metodai skirstomi į 3 pagrindines klases:

- mokymas su mokytoju (angl. *supervised learning technique*);
- mokymas be mokytojo (angl. *unsupervised learning technique*);
- kita.

Į pirmą kategoriją „mokymas su mokytoju" patenka klasifikacijos ir prognozavimo uždaviniai. Antrai kategorijai „mokymas be mokytojo" priskiriamas klasterizavimo ir asociacinių taisyklių paieškos uždavinys. Vizualizavimas, nuokrypių identifikavimas ir ryšių analizė neskirstomi į „mokymo su mokytoju" arba „mokymo be mokytojo" klases.

DT užduotims spręsti parenkami tinkami algoritmai. Tiek DT metodo parinkimas, tiek optimalaus algoritmo parametrizavimas priklausys nuo suformuotos analizės užduoties tikslų ir turimų duomenų charakteristikų.

Per pastarąjį dešimtmetį sukaupta tam tikra DT metodų taikymo medicinoje patirtis. Diagnostikoje plačiai taikomi neuroniniai tinklai, sprendimų medžiai, sprendimų taisyklės [8], asociacinių taisyklių paieškos metodai pritaikomi kaštų analizei [7], paciento būsenai ir pasveikimo tikimybei prognozuoti, plačiai naudojamos įvairių prognozavimo algoritmų kombinacijos [4].

Žemiau išvardyti populiariausi duomenų tyrybos metodai ir technikos pagal *KDnuggets* [9] atliktą apklausą:

1. Sprendimų medžiai ir sprendimų taisyklės;
2. Regresija;
3. Klasterizacija;
4. Priklausomybes tirianti aprašomoji statistika;
5. Vizualizavimas;
6. Laiko eilučių analizė;
7. Atraminių vektorių klasifikatorius;
8. Asociacinių taisyklių paieškos;
9. Sudėtiniai metodai;
10. Teksto tyryba.

2014 m. paskelbtame N. Esfandiari, M. R. Babavalian ir kt. darbe [1] sistemiškai apžvelgiama literatūra, kurioje aprašomi DT taikymai medicinoje sustruktūrintiems duomenims analizuoti. Anot autorių, medicinoje populiariausi DT klasifikavimo (sprendimų medžiai, neuroniniai tinklai, sprendimų taisyklės, atraminių vektorių modelis) ir klasterizavimo (k-vidurkių ir hierarchinis klasterizavimas) metodai bei asociacijų paieška (*a priori* asociacijos taisyklių paieška).

Trumpai aprašysime kelis populiariausius DT metodus.

Klasifikavimas naudojamas siekiant objektus priskirti iš anksto numatytoms klasėms. Klasės vaidmenį atlieka pasirinktas atributas duomenų rinkinyje, determinuojantis objektą. Statistikoje toks atributas vadinamas priklausomu kintamuoju. Klasifikuojant objektus algoritmas sukuria klasifikavimo modelį, kuris gali būti toliau pritaikytas naujiems duomenims. Pavyzdžiui, klasifikavimo algoritmo sukurtas diagnostinis krūties vėžio modelis toliau gali būti pritaikytas sprendimo priėmimo palaikymo sistemoje diagnozuojant pacientą, kurio duomenys nebuvo naudojami prognozavimo modeliui kurti. Klasifikavimas yra dviejų žingsnių procesas, susidedantis iš mokymo ir testavimo. Mokymo žingsnio metu algoritmas analizuoja mokymuisi skirtus duomenis ir sukuria klasifikavimo modelį. Testavimo metu modelio tikslumas tikrinamas su kitu duomenų rinkiniu. Populiariausi klasifikavimo metodai: sprendimų medžiai, Bayeso klasifikatoriai, dirbtiniai neuroniniai tinklai.

Klasterizavimas apibrėžiamas kaip mokymas be mokytojo. Tai reiškia, kad klasterizuojant nereikalingas *a priori* žinojimas, kokiai grupei (klasteriui) priklauso objektas. Klasterizavimo algoritmas, taikydamas euristinius metodus, suskirsto objektus į numatytą grupių skaičių pagal jų duomenų panašumą. Panašumo matmuo gali būti parinktas atsižvelgiant į objektą apibūdinančius atributus. Objektų panašumui įvertinti dažnai naudojamos distancijos metrikos: Euklido, Manhattano, Jackkardo ir kt.

Labiausiai paplitę klasterizavimo metodai: hierarchinis ir particijų klasterizavimas.

Asociacinių taisyklių paieškos metodą 1991 m. pasiūlė Piatetsky-Shapiro. Šis kitaip dar vadinamas pirkėjo krepšelio analizės metodas padeda rasti netrivialius dėsningumus duomenyse. Asociacinės taisyklės nusako ryšius tarp duomenų elementų. Tipinis pavyzdys: analizuojant pirkėjų krepšelius nustatytas dėsningumas, kad pirkėjai, nusipirkę duonos ir sviesto, taip pat perka ir pieno. Labiausiai paplitęs *a priori* algoritmas numato du įvesties parametrus: taisyklių palaikymą ir stiprumą.

Asociacinės taisyklės stiprumas yra procentinė duomenų aibės dalis, kuriai ši taisyklė galioja. Pavyzdžiui, 80 proc. taisyklės stiprumas reikštų, kad 80 proc. pirkėjų, pirkusių duonos ir sviesto, nusipirko

pieno. Asociacinės taisyklės palaikymas yra procentinė duomenų aibės dalis, kurioje randama taisyklės sąlyga. Pavyzdžiui, 20 proc. taisyklės palaikymas reikštų, kad iš viso 20 proc. pirkėjų pirko duonos ir sviesto.

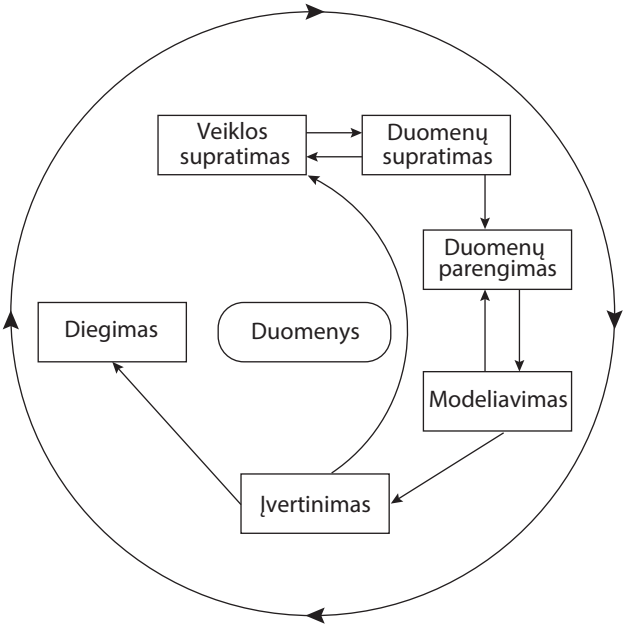
Duomenų tyryba kaip žinių gavimo proceso dalis

Duomenų tyrybos metodai pagal paskirtį skirstomi į dvi grupes: prognozavimo ir duomenų apibūdinimo. Apibūdinančių uždavinių tikslas – rasti interpretuojamus dėsningumus (angl. *patterns*) ir tarpusavio ryšius, o prognozavimo uždaviniai siekia išpranašauti tam tikrus įvykius arba tam tikras nežinomas reikšmes nagrinėjamoje interesų sferoje. Pagrindinis metodinis skirtumas yra tai, kad prognozavimas reikalauja į pirminius duomenis įtraukti specialų prognozuojamąjį kintamąjį (klasę). Atsakymas gali būti skaitinis arba kategorinis, atitinkamai duomenų tyrybos metodai, kurie skirti prognozavimui, skirstomi į regresijos arba klasifikavimo [4].

Nagrinėtose publikacijose duomenų tyrybos taikymą medicinoje siūloma remti bendrai priimtomis duomenų tyrybos metodikomis. Intelektualios duomenų analizės procesas naudojant duomenų tyrybos ir statistinės analizės metodus yra iteracinis. Literatūroje siūlomi įvairūs DT proceso modeliai: CRISP-DM, Simoudis 1996 m. proceso modelis, Fayyad ir kt. 1996 m. proceso modelis, Cabena 1998 m. proceso modelis, Cios 2000 m. proceso modelis [10–12, 21].

Plačiausiai paplitusi duomenų tyrybos proceso metodika – *Cross-Industry Standard Process for the Data Mining* (CRISP-DM) [9]. CRISP-DM apibrėžia procesų modelį [21], kuris duomenų tyrybą išskaido į šešias fazes: veiklos supratimą, duomenų supratimą, duomenų parengimą, modeliavimą, įvertinimą ir diegimą (1 pav.). Metodikoje numatyta kiekvienos fazės įvestis, išvestis ir vykdymo strategija.

CRISP-DM traktuoja duomenų tyrybos procesą kaip klasikinį projektą, kuris turi apibrėžtą tikslą bei pagrindinius projekto apribojimus – laiką, išteklius ir apimtį. Kadangi projektas turi jau numatytą tikslą, CRISP-DM neakcentuoja užduoties suformulavimo. Tačiau, kaip pabrėžia P. Baylis [13], duomenų tyryba medicinoje prasideda būtent nuo teisingo užduoties suformulavimo, kai klinicistai kartu su duomenų analizės specialistais numato probleminę sritį ir, analizuodami veiklos sritį bei SIS prieinamus duomenis, suformuluoja problemą ir techninę užduotį.



1 pav. CRISP DM procesų modelis

Veiklos supratimas

Pradinėje fazėje akcentuojamas duomenų analizės projekto tikslų supratimas ir reikalavimų iš dalykinės srities perspektyvos, taip pat šios suformuluotos problemos vertimas į duomenų tyrybos problemos apibrėžimą. Šioje fazėje nustatomas preliminarus planas, kaip bus siekiama tikslo.

Duomenų supratimas

Duomenų supratimo fazė prasideda nuo pradinių duomenų rinkimo ir pažinties su duomenų rinkiniu. Turi būti nustatytos duomenų kokybės problemos, suformuluotos pirmos prielaidos, kokie duomenų rinkiniai gali būti įdomūs tyrybai.

Duomenų parengimas

Duomenų parengimo fazė apima visas veiklas, reikalingas galutiniam duomenų rinkiniui parengti. Šios fazės veiksmai labai priklausys nuo turimų pradinių neapdorotų duomenų. Tipiniai duomenų parengimo uždaviniai yra lentelės, įrašo ir atributo projekcijų parinkimas, atributų transformacija, normalizavimas, kategorizavimas, triukšmų šalinimas, diskretizavimas.

Modeliavimas

Šioje fazėje parenkami tinkami modeliavimo metodai, algoritmai arba jų deriniai. Toliau parenkamos optimalios algoritmų parametrų reikšmės. Dažniausiai galimi keli modeliavimo metodai tam pačiam

uždaviniui išspręsti. Kai kurie modeliavimo metodai turi specialius reikalavimus duomenims. Dėl to dažnai šis etapas vykdomas iteraciniu būdu, kol pasiekiamas užsibrėžtas modelio kokybės kriterijus.

V. Špečkauskienė ir A. Lukoševičius [11] siūlo 11 žingsnių metodiką optimaliam DT algoritmui parinkti:

1. Surinkti ir susipažinti su keliais klasifikavimo algoritmais;
2. Analizuoti duomenų rinkinį;
3. Atrinkti duomenų rinkiniui tinkamus algoritmus;
4. Testuoti visą duomenų rinkinį su atrinktais klasifikavimo algoritmais ir standartinėmis parametrų reikšmėmis;
5. Pasirinkti tolesnei analizei geriausius algoritmus;
6. Apmokyti atrinktus algoritmus su sumažintu duomenų rinkiniu, eliminuojant atributus, kurie pasirodė neinformatyvūs konstruojant ir vizualizuojant sprendimų medžius;
7. Naudojant optimalų duomenų rinkinį, suformuotą kiekvienam algoritmui iš naudingiausių duomenų, identifikuotų 6 žingsnyje, koreguoti standartines algoritmų parametrų reikšmes;
8. Įvertinti rezultatus;
9. Sumaišyti duomenų rinkinio atributų reikšmes atsitiktine tvarka;
10. Atlikti 6 ir 7 žingsnius su nauju duomenų rinkiniu;
11. Įvertinti ir palyginti rezultatus ir algoritmų našumą.

Pasiūlyta metodika gali būti iš dalies automatizuojama naudojant specializuotą programinę įrangą, kurią siūlo autoriai.

Įvertinimas

Šioje fazėje jau suformuotas aukštos kokybės modelis (arba keli modeliai). Prieš galutinai įdiegiant modelį labai svarbu nuodugniai jį įvertinti, peržiūrėti modelio konstravimo žingsnius, įsitikinti, kad veiklos tikslai yra pasiekti tinkamai. Formaliai įvertinama modelio kokybė. Modelio kokybei įvertinti naudojamos DT ir statistikoje populiarios metrikos: bendras tikslumas, jautrumas, specifiškumas, ROC kreivė, teigiamas ir neigiamas tikimybinis santykis, teigiama ir neigiama prognostinė vertė. Bendras tikslumas – procentas teisingai klasifikuotų objektų. Jautrumas – teisingai klasifikuotų teigiamų rezultatų dalis iš rezultatų visumos. Specifiškumas – teisingai klasifikuotų neigiamų tyrimų rezultatų dalis iš rezultatų visumos. Jautrumo ir specifiškumo sąryšį galima įvertinti taikant ROC kreivę (angl. *Receiver Operating Characteristic*) arba ploto po kreive skaitinę išraišką (angl. *area under curve*, AUC).

Galutinis šios fazės rezultatas – sprendimas, ar duomenų tyrybos rezultatai gali būti naudojami.

Diegimas, arba naujų žinių ir duomenų tyrybos rezultatų panaudojimas

Modelio sukūrimas nėra duomenų tyrybos projekto pabaiga. Netgi tuo atveju, jei duomenų tyrybos projekto tikslas buvo sužinoti daugiau apie turimus duomenis, gautos žinios turi būti organizuotos ir pristatytos galutiniam vartotojui suprantama forma. Priklausomai nuo reikalavimų, paprasčiausiu atveju diegimo fazę gali sudaryti ataskaitos parengimas arba pasikartojančio duomenų tyrybos proceso įdiegimas. Panaudojant PMML modeliavimo kalbą, gautas prognozavimo modelis gali būti išsaugotas ir eksportuotas tolesniam naudojimui sveikatos priežiūros įstaigų klinikinių arba valdymo sprendimų palaikymo sistemose. Dažnai galutinis vartotojas, o ne duomenų analitikas vykdys diegimo veiklas. Labai svarbu, kad galutinis vartotojas iš anksto numatytų, kokie veiksmai turės būti atlikti siekiant gauti praktinę naudą iš sukurto DT modelio.

Sprendžiant sveikatos priežiūros specialistų suformuotas užduotis, pagrindinis duomenų tyrybos specialisto uždavinys – rasti ir pritaikyti tinkamus DT metodus, galinčius nustatyti atributų sąryšius ir suformuoti tinkamą modelį.

Duomenų tyryba turi platų instrumentų spektrą ir keli skirtingi metodai gali vienodai gerai tikti tam pačiam tikslui pasiekti. Dėl šios priežasties gali būti nepraktiška nagrinėti visus alternatyvius metodus, ir konkretaus metodo pasirinkimą nulemia ne tik objektyvios analizės rezultatai, bet, kaip teigia R. Bellazzi ir kt. [4], ir duomenų tyrybos eksperto intuicija.

Duomenų tyrybos taikymas sveikatos srityje

Prielaidos ir priešistorė

Medicininės informacijos rinkimas ir jos rutininė statistinė analizė vykdoma nuo viduramžių. Pirmas žinomas medicininis leidinys, analizuojantis medicininę statistinę informaciją, išleistas 1662 m. Londone [14]. 1863 m. šiuolaikinės medicininės slaugos pradininkė F. Nightingale savo užrašuose nuogąstavo dėl sveikatos įrašų trūkumo bei nesistemingo saugojimo ligoninėse, nes tai neleisdavo analizuoti kaštų ir gydymo efektyvumo. 1977 m. JAV kongresas išleido studiją „Medicininių informacinių sistemų praktikų pasekmės" [15]. Joje teigiama, kad sveikatos informacinės sistemos gali būti naudojamos mokymams, padėti medicinos ir sveikatos priežiūros specialistams teikiant aukštesnės kokybės paslaugas pacientams ir

optimizuojant SPĮ veiklą. Studijos autoriai teigė, kad galiausiai tokios sistemos pateiks duomenis ir žinias, kurios anksčiau buvo neprieinamos mokslininkams ir sveikatos apsaugą reglamentuojančioms institucijoms. Nuo 2000 m. visame pasaulyje aktyviai diegiamos regioninės ir nacionalinės elektroninės sveikatos istorijos sistemos, kurių tikslas – kaupti visus reikšmingus paciento medicininius dokumentus.

Kaupiamos medicininės informacijos svarba įvertinta ir įprasminta prieš kelis šimtmečius. Kai pirminiai duomenys kaupiami el. būdu, be standartinių sveikatos priežiūros srityje priimtų statistinių rodiklių skaičiavimo, atsiranda galimybė taikyti duomenų tyrybos metodus.

Tikslai ir uždaviniai

Anot P. Baylis [13], raktas į sėkmingą medicininių duomenų tyrybą yra teisingas SPĮ veiklos ar klinikinės problemos nustatymas. DT metodai dažniausiai atlieka biomedicininių duomenų regresijos, klasterizavimo, klasifikacijos ir vizualizavimo uždavinius, siekiant palengvinti sprendimų priėmimą sveikatos priežiūros specialistams [16]. Pasak N. Esfandiari ir kt., galime išskirti keturis pagrindinius DT taikymo tikslus medicinoje:

- efektyvumo didinimas ir žmogiškojo faktoriaus eliminavimas. Sprendžiami tam tikrų ligų diagnozavimo uždaviniai, kai itin svarbus tikslumas;
- laiko ir sąnaudų mažinimas. Taikoma, kai įprasti diagnozavimo metodai užima daug laiko arba yra labai brangūs;
- medicininių sprendimų palaikymo sistema. Naudoja kelių procesų automatizaciją, pvz., prognozavimo modelius ir ekspertines sistemas. Gali būti taikoma kaip pagalba mažiau patyrusiam ar žemesnės kvalifikacijos medicinos personalui;
- žinių gavimas. Naudojama naujoms žinioms gauti arba hipotezėms formuluoti.

Apibendrinant publikuotuose darbuose sprendžiamus DT taikymo medicinoje uždavinius galima teigti, kad jie priskiriami prie gydymo išteklių optimizavimo arba gydymo kokybės gerinimo tikslų. 1 lentelėje pateikiami atitinkamai sustruktūrinti DT medicinoje tikslai ir uždaviniai.

Duomenų tyrybos medicinoje problematika

Šiame skyriuje pateikiamos specifinės medicinos taikomosios srities duomenų tyrybos problemos. Jų visuma leidžia pamatyti tik šiai sričiai būdingas problemas, kurios atskiria duomenų analizės ir tyrybos procesą medicinoje nuo duomenų analizės ir tyrybos kitose dalykinėse srityse.

1 lentelė. Duomenų tyrybos medicinoje tikslai ir uždaviniai

Tikslai	Uždaviniai
Gydymo išteklių optimizavimas	• Potencialių kaštų mažinimo ir pajamų didinimo galimybių identifikavimas [7] • Paciento stacionarizavimo trukmės priklausomybės[5] nuo paciento demografinių duomenų, anamnezės, pasirinkto gydymo metodo ir kitų faktorių nustatymas [13] • Rehospitalizacijos tikimybės prognozavimas • Pooperacinių komplikacijų bei jų tikimybių prognozavimas • Medicinos personalo efektyvumo rodiklių prognozavimas • Netikslingų medicininių nurodymų nustatymas • Netinkamų vaistų paskyrimų nustatymas • Ankstyva ligų diagnostika (*screening*)
Gydymo kokybės gerinimas	• Ankstyva ligų diagnostika (*screening*) • Komplikacijų tikimybės įvertinimas • Ligos eigos modeliavimas [20] • Specifinių klinikinių atributų asociacijų nustatymas, siekiant tikslinti diagnozę arba parinkti gydymo planą • Realiu laiku fiksuojamų daugiamačių biomedicininių duomenų apibendrinimas, siekiant palengvinti sprendimų priėmimą [5] • Biomedicininių duomenų rinkinių kokybės analizė [4]: - duomenų rinkinio pilnumo nustatymas - duomenų rinkinio fragmentiškumo nustatymas • Diagnozės nustatymas arba diagnozės tikslinimas • Medicininių ekspertinių sistemų žinių bazės formavimas • Medikamentų efektyvumo prognozavimas • Mikromasyvų analizė sprendžiant uždavinius: - ankstyvoji ligų diagnostika - individualių gydymo būdų parinkimas - ligos pasireiškimo tikimybės nustatymas

Jeigu DT procesas būtų paprastas, informacijos vadybos problemos būtų seniai išspręstos (R. Bellazzi, B. Zupan). Kaip pabrėžiama daugelyje šaltinių [7, 11, 12, 16], praktinis DT pritaikymas medicinoje turi nemažai kliūčių: technologinių, tarpdisciplininės komunikacijos, etikos ir paciento duomenų apsaugos. Be to, yra keletas gerai žinomų biomedicininių duomenų problemų, kaip netiksli ir fragmentuota informacija. Netikslios informacijos pavyzdžiai: gyvybinių funkcijų matavimai atliekami pacientui esant neramios būsenos; tyrimams reikalingas pasėlis paimtas nesteriliomis sąlygomis; matavimo prietaisų paklaidos. Fragmentuotos informacijos pavyzdžiai: paciento EKG duomenų masyvai kaupiami tik po miokardo infarkto, nėra visos paciento tyrimų informacijos.

Kitas duomenų tyrybos medicinoje bruožas yra jos rezultatų panaudojimas priimant žmogaus gyvybei

[5] LOS, angl. *length of stay.*

kritiškus sprendimus. Dėl to, kaip nurodo S. Laxman ir P. S. Sastry, pasirinkto DT metodo rezultatai turi būti deskriptyvūs, t. y. pateikti su paaiškinimais taip, kad medicinos ekspertai galėtų suprasti, kaip gauti šie rezultatai. Dėl to galima teigti, kad vieni DT metodai, pavyzdžiui, sprendimų medžiai, šiuo požiūriu labiau tinka nei neuroniniai tinklai.

Kelių medicinos specialybių duomenų analizė iškelia papildomus uždavinius. Medicinoje semantiškai ta pati koncepcija gali turėti kelis pavadinimus bei jiems priskirtus skirtingus kodus iš skirtingų kodifikatorių. Panagrinėkime hipotetinį pavyzdį. Ligoninės *X* patologoanatomijos departamente pritaikyta klinikinė patologoanatomijos sistema, kuri naudoja SNOMED CT ontologijos žodyną. Tos pačios ligoninės kardiologijos departamentas naudojasi kardiologų informacine sistema, kurioje įdiegti TLK-10, TLK PT klasifikatoriai, praplėsti pagal kardiologų poreikius. Taip pat kardiologai yra radiologinių tyrimų naudotojai. Radiologiniai tyrimai yra kompiuterizuoti, tačiau neintegruoti. Radiologinė informacija saugoma DICOM formatu ir naudojamas TLK-10 ligų klasifikatorius. Iškeliamas prognozavimo uždavinys, kuriam išspręsti reikalinga pacientų klinikinė informacija, sujungianti kardiologijos ir patologoanatomijos departamentuose sukauptą informaciją. Taip pat tyrimui svarbūs radiologinių tyrimų duomenys. Šioje vietoje susiduriame su kaupiamos informacijos sąsajumo problema. Prieš taikant duomenų tyrybos algoritmus, duomenys turės būti homogenizuojami pasitelkiant atitikimo nustatymo[6] metodus. Norint pasinaudoti radiologijos tyrimų rezultatų duomenimis, teks taikyti kompiuterizuoto vaizdų apdorojimo algoritmus arba daryti semantinę tekstinių tyrimų aprašymų analizę – tekstinę tyrybą.

Tais atvejais, kai informacinėse sistemose naudojami standartizuoti biomedicininiai klasifikatoriai, nomenklatūriniai sąrašai ar ontologijos, sąsajumo nustatymo uždavinys – teisingos bendros ontologijos parinkimas. Tačiau kartais ir tai bus neįmanoma, nes įstaigos naudoja klasifikatorių ir nomenklatūrų plėtinius arba nacionalines versijas, kurios netapačios tarptautinėms versijoms. Tokiais atvejais duomenų tyrybos ir medicinos informatikos specialistai turi sukurti duomenų transformacijos metodus duomenų sąsajumui užtikrinti.

Sprendžiant ligoninės *X* pradinių duomenų homogenizacijos uždavinį, reikės spręsti ir **medicininių informacinių sistemų sąveikumo** (angl. *interoperability*) **problemą**. Kadangi kardiologų, patologoanatomų ir radiologų informacinės sistemos nėra integruotos, reikės užtikrinti šių sistemų duomenų sujungimą.

Medicinos informatika siūlo keletą sąveikumo standartų. Lietuvos e. sveikatos strategijoje numatytas HL7 3 versijos paciento administracinių ir klinikinių duomenų apsikeitimo standartų platformos panaudojimas. Tačiau Lietuvos praktikoje naudojamų šio standarto taikymų dar nėra. SIS sąveikumo praktinis įgyvendinimas priklausys nuo konkrečių SIS palaikomų standartų. Idealiu atveju visos naujos kartos SIS turėtų palaikyti industrinius duomenų apsikeitimo standartus (HL7, HL7 CDA, DICOM) ir naudoti tarptautinius klasifikatorius. Praktikoje situacija priešinga. Todėl sėkmingam DT metodų pritaikymui keliamas dar vienas papildomas uždavinys – informacinių sistemų integracija. Čia sistemų integracija turi būti suprantama plačiąja prasme, pradedant duomenų apsikeitimo architektūra ir baigiant semantiniu duomenų integralumu.

Sąveikumo problematika tampa dar sudėtingesnė integruojant skirtingų valstybių SIS. Medicinos informatikoje sukurta nemažai persidengiančių standartų, iš jų vieni labiau paplitę Europoje, kiti – JAV ar Australijoje. Tačiau tarptautinė medicininės informatikos bendruomenė sprendžia šią problemą [17].

Kita dažna biomedicininės informacijos analizę ir tyrybą nagrinėjančių publikacijų tema – **etika arba paciento duomenų konfidencialumas**. Dažniausiai įstatymai, saugodami asmens privatumą, draudžia naudotis paciento klinikine informacija be paciento sutikimo. Tai apsunkina klinikinės informacijos panaudojimą moksliniais tikslais. Ši problema sprendžiama depersonalizuojant duomenis. Tai atliekama atskiriant pacientą identifikuojančią asmeninę informaciją nuo kitos demografinės informacijos. Į tyrimams naudojamą duomenų masyvą neturi patekti paciento vardas, pavardė, paso arba draudimo numeriai, kiti asmenį identifikuojantys atributai. Vietoje paciento asmenį identifikuojančių atributų naudojami sintetiniai identifikatoriai, kurie negali būti panaudoti asmens tapatybei nustatyti.

Tam tikrose šalyse, kur stipri lygių galimybių įstatyminė bazė, etikos problematika neapsiriboja duomenų depersonalizavimu. Pavyzdžiui, JAV sprendimai dėl paslaugų teikimo negali būti priimami vadovaujantis rasės, lyties arba amžiaus kriterijumi. Kadangi šie demografiniai paciento rodikliai yra labai svarbūs ir dažnai naudojami SIS DT, iškyla keblumų dėl vykdomų analizių rezultatų panaudojimo gydymo tikslais.

[6] Angl. *mapping*.

Nekorektiškos ir fragmentuotos informacijos problemos

Konstruojant DT modelius ir parenkant veiksmingus algoritmus, kurių aukštas specifiškumas ir jautrumas, visada turi būti įvertinamas klinikinės informacijos patikimumas bei išsamumas, kitaip sakant, duomenų kokybė. Duomenų korektiškumui įtaką gali padaryti neteisingi matavimai, nesuderinta laboratorinė įranga, paciento apžiūra nepalankiomis sąlygomis. Dėl šių priežasčių, parenkant analizuojamų duomenų rinkinį, turi būti imama kuo didesnė klinikinių duomenų imtis. Tokiu būdu ignoruojant netipines duomenų išskirtis sumažinama nekorektiškos informacijos įtaka analizės rezultatams.

Sveikatos informacinėse sistemose kaupiama informacija

Nuo XX a. pr. ligoninėse vykdomas rutininis veiklos rodiklių stebėjimas ir fiksavimas. Tarptautinės ir nacionalinės sveikatos priežiūros organizacijos skelbia rodiklius, kuriais remiantis galima įvertinti ir palyginti priešoperacinį stacionarizavimo laiką, lovų apyvartą, letališkumą, pacientų apsilankymų srautą ir kt. Šiems rodikliams skaičiuoti būtini duomenys renkami popierine arba elektronine forma. El. duomenų kaupimo atveju SPĮ gali analizuoti kaupiamą informaciją realiu laiku ir priimti atitinkamus valdymo sprendimus. Tokiu būdu vienas iš informacijos perkėlimo į elektroninę terpę tikslų – palaikyti SPĮ valdymo sprendimus.

Pavyzdžiui, JAV ligoninėse „Kaiser Permanente" nuolatos matuojami tiek paciento sveikatos, tiek medicinos personalo darbo rodikliai, kurie toliau naudojami gydymo procesui gerinti, medicinos personalo darbui įvertinti ir lyginamajai analizei bei mokslininiams tyrimams.

Šiuo metu medicinoje taikoma daugybė skirtingos paskirties informacijos klasifikavimo sistemų. Medicininių ontologijų tiriamąjį darbą vykdo kelios pasaulinės organizacijos ir universitetai. Nemažai dėmesio skiriama esamų formalių, pusiau formalių ir neformalių ontologijų semantinei analizei ir jų integralumo tyrimui [18]. Esminės sprendžiamos problemos – medicininių ekspertinių sistemų kūrimas ir įvairių sričių SIS integralumo užtikrinimas dėl naudojamų skirtingų klasifikacijos sistemų. Sprendžiant šiuos uždavinius nagrinėjamas esamų medicininių terminų žodynų sąsajumas, vykdomas ontologijų formalizavimas.

Bet kuri ontologija būtinai privalo turėti terminų žodyną ir jų reikšmių specifikaciją. Tai apima apibrėžimus ir sąvokų tarpusavio ryšius, kurie apibrėžia dalykinės srities struktūrą ir neleidžia neteisingai interpretuoti terminus.

Biomedicininių DT ontologijos naudojamos sprendžiant šiuos uždavinius [19]:

- medicininių terminų normalizavimas;
- skirtingų klasifikacijų, žodynų ar nomenklatūrų integralumas;
- informacijos abstrakcijos lygio kėlimas.

Taip pat DT metodai naudojami ontologijoms kurti.

O. Bodenreider nurodo [19] šias dažniausiai pasaulyje naudojamas biomedicinines ontologijas: ICD, LOINC, SNOMED, FMA, GO, *RxNorm*, *MeSH*, NCI *Thesaurus*, UMLS. 2 lentelėje nurodytas išvardytų žinynų, klasifikatorių, nomenklatūrų ir ontologijų konceptų skaičius.

2 lentelė. Populiariausi biomedicininiai terminų žinynai ir ontologijos

Pavadinimas	Konceptų skaičius
Tarptautinis ligų klasifikatorius (*International Classification of Diseases*, ICD)	12 318
Laboratorinių tyrimų klasifikatorius (*Logical Observation Identifiers, Names and Codes*, LOINC)	46 406
Klinikinių nomenklatūrų sąrašai (SNOMED *Clinical Terms*, SNOMED CT)	310 314
Anatominis modelis (*Foundational Model of Atonomy*, FMA)	~72 000
Genų ontologija (*Gene Ontology*, GO)	22 546
Medikamentų žinynas (*RxNorm*)	93 426
Medicininių temų antraščių repozitorijus (*Medical Subject Headings*, MeSH)	24 767
Nacionalinio vėžio instituto enciklopedija (NCI *Thesaurus*)	58 868
Unifikuota medicininės kalbos sistema (*Unified Medical Language System*, UMLS)	1,4 M

JAV Nacionalinė medicinos biblioteka nuo 1986 m. kuria integralią biomedicininę ontologiją UMLS (angl. *Unified medical language system* – unifikuota medicininės kalbos sistema).

UMLS yra apibendrinanti biomedicininė ontologija, kuri leidžia sveikatos priežiūros specialistams ir mokslininkams išrinkti ir integruoti informaciją iš skirtingų elektroninių biomedicininių šaltinių, pradedant paciento elektroninių sveikatos įrašų sistemomis, baigiant žinių bazėmis. Vienuoliktoje UMLS versijoje yra daugiau kaip 1,4 mln. koncepcijų ir 6 mln. sąryšių. Koncepcijose unifikuoti sinonimai iš 100 skirtingų klasifikacijos sistemų, tokių kaip MeSH, ICD-10, ICD-9-CM, SNOMED.

UMLS sudaro šie komponentai:

- metažinynas (*metathesaurus*) – UMLS duomenų bazė, kurią sudaro koncepcijų ir terminų žodynas,

jų tarpusavio sąrašai bei nuorodos į išorinius žodynus (klasifikatorius);
- semantinis tinklas – kategorijų ir jų sąryšių, kurie naudojami esybių apibrėžimuose metažinyne, rinkinys;
- specialisto žodynas – leksikografinis žodynas, naudojamas natūralios kalbos apdorojimui;
- programinės įrangos rinkinys.

UMLS ypač naudinga, kai integruojant įvairius duomenų šaltinius reikia spręsti vienos koncepcijos skirtingų išraiškų problemą. Pavyzdžiu galima pateikti vieną ligos konceptą – „Adisono liga". Šis konceptas priklausomai nuo klasifikavimo sistemos taip pat gali būti vadinamas: „Pirminis hiperadrenalizmas", „Pirminis antinksčių žievės nepakankamumas", E27.1 ir t. t.

Visi pavadinimai ir kodai pavaizduoti 2 pav. ir yra visiškai arba iš dalies tapatūs konceptui „Adisono liga".

Adisono liga	MeSH	D000224
Pirminis antinksčių žievės nepakankamumas	TLK-10	E27.1
Pirminis adrenalitas	MedDRA	10036696
Adisono liga (sutrikimas)	SNOMED CT	363732003
C0001403	**Adisono liga**	

2 pav. UMLS koncepcijos „Adisono liga" sąsajumo pavyzdys

Šiuo metu Lietuvoje įdiegti nacionaliniai medicininiai registrai, kurie įgalintų standartizuotą duomenų apsikeitimą unifikuotais formatais. Valstybinių ligonių kasų administruojamose nacionalinėse informacinėse sistemose yra valdomi ir SPĮ naudoti teikiami šie klasifikatoriai:
- TLK-10-AM sisteminis ligų sąrašas;
- ACHI sisteminis intervencijų sąrašas;
- Didžiųjų ir mažųjų operacijų sąrašas;
- Gimininių diagnozių grupių sąrašas;
- Socialinio draudimo kompensuojamų paslaugų klasifikatorius;
- Lietuvos gydytojų klasifikatorius;
- Lietuvos sveikatos priežiūros įstaigų klasifikatorius.

Sėkmingam duomenų perdavimui tarp įvairių informacinių sistemų naudojami medicininių duomenų perdavimo standartai. HL7[7] yra standartų sistema, apibrėžianti klinikinių ir administracinių pacientų duomenų apsikeitimą ir integravimą. HL7 2-os versijos standartas, kuris labiausiai paplitęs pasaulyje, įgalina el. žinučių būdu keistis paciento administracine, finansine ir klinikine informacija. Kadangi 2-os versijos protokolas pradėtas kurti prieš 20 metų, jo sintaksė nėra pritaikyta naudoti šiuolaikines į paslaugas orientuotas sistemas[8]. 2005 m. buvo paskelbta pradinė HL7 3 versija, kuri pagrįsta formalia metodologija HDF ir remiasi objektiškai orientuota paradigma. Lietuvos nacionalinė e. sveikatos sistema duomenims keistis naudoja HL7 3 versiją. Kitas svarbus klinikinių duomenų kaupimo ir perdavimo standartas, įeinantis į HL7 3 v., – HL7 CDA[9]. HL7 klinikinių dokumentų architektūra yra XML kalbos sintaksės standartas, nurodantis klinikinių dokumentų kodavimą, struktūrą ir semantiką, užtikrinantis unifikuotą klinikinių dokumentų struktūrą. HL7 CDA standartas išsprendžia klinikinių dokumentų apsikeitimo tarp įvairių lokalių ir nacionalinių informacinių sistemų problemą.

APIBENDRINIMAS

Visuomenės sveikatos informacijos rinkimas ir jos rutininė statistinė analizė vykdoma nuo viduramžių. Pirmas žinomas medicininis leidinys „London Bills of Mortality", analizuojantis medicininę statistinę informaciją, buvo išleistas 1662 m. Londone. Šiuolaikinė duomenų analizė medicinoje sujungia duomenų tyrybą ir klasikinę statistinę analizę. Tačiau duomenų tyrybos metodai medicinoje naudojami tik kelis dešimtmečius ir, galima sakyti, yra paauglystės amžiaus.

Dauguma Lietuvos SPĮ iš dalies kompiuterizuotai renka medicininius duomenis. Visa tai sudaro labai palankią erdvę DT taikyti Lietuvos sveikatos priežiūros institucijose. Sveikatos informacinėse sistemose duomenys išsaugomi dideliais kiekiais, kurie agreguojami iš įvairių šaltinių ir kurių kokybė bei struktūra yra nevienoda.

DT taikymas medicinoje nuo kitų sričių skiriasi tuo, kad pradiniai duomenys yra heterogeniški, klinikinių duomenų panaudojimas DT gali būti ribojamas etikos, socialiniais bei teisiniais aspektais. Šie apribojimai DT susiję su privatumo ir duomenų saugumo grėsmėmis, pacientų teisinių ieškinių galimybe ir būtinybe įvertinti DT privalumus bei potencialią klaidingų sprendimų riziką.

Straipsnyje apžvelgtos DT taikymo medicinoje problemos:

[7] HL7, angl. *Health level 7*.

[8] SOA, angl. *service oriented architecture*.

[9] CDA, angl. *clinical document architecture*.

- pirminių duomenų surinkimo problema:
  - sveikatos informacinių sistemų sąveikumas,
  - pirminių duomenų homogenizavimas (bendros ontologijos parinkimas),
  - pacientų duomenų apsaugos užtikrinimas (pacientų duomenų depersonalizavimas);
- pirminių duomenų tyrimams kokybės užtikrinimas:
  - duomenų pilnumo analizė,
  - duomenų patikimumo analizė,
  - duomenų validumo analizė.

Taikant DT metodus turi būti sprendžiama keletas multidisciplininių uždavinių: teisingo klinikinio arba vadybos uždavinio suformulavimas, pirminių duomenų parengimas, deskriptyvaus DT metodo parinkimas ir jo pritaikymas uždaviniui.

Atkreiptinas dėmesys į DT traktuotę, kaip žinių gavimo proceso dalį. DT taikymo metodika CRISP-DM detaliai aprašo visą žinių gavimo proceso modelį, pradedant tikslų suformulavimu ir baigiant modelio diegimu organizacijoje. Industrinių procesų modelių pritaikymas užtikrina aukštesnę DT tiriamųjų projektų kokybę ir jų rezultatų panaudojimą praktikoje.

Straipsnis gautas 2014-10-27, priimtas 2014-11-28

Literatūra

1. Esfandiari N, Babavalian MR, Moghadam AM, Tabar VK. Knowledge discovery in medicine: Current issue and future trend. Expert Systems with Applications, Elsevier. 2014;44:4434-4463
2. Berka P, Rauch J, Zighed DA. Data Mining and Medical Knowledge Management – Cases and Applications. Idea Group Inc (IGI). 2009;440.
3. Stühlinger W, Hogl O, Stoyan H, Müller M. Intelligent Data Mining for Medical Quality Management. 14th European Conference Artificial Intelligence; 2006.
4. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: Current issues and guidelines. International journal of medical informatics. 2008;77:81-97.
5. Stacey M, McGregor C. Temporal abstraction in intelligent clinical data analysis: A survey. Artificial Intelligence in Medicine. 2007;39:1-24.
6. Standarto aprašymas (Predictive Model Markup Language). Prieiga per internetą: <http://www.dmg.org/v4-0/GeneralStructure.html> [žiūrėta 2012-08-10].
7. Silver M, Sakata T, Su HC, Herman C, Dolins SB, O'Shea MJ. Case Study: How to Apply Data Mining Techniques in a Healthcare Data Warehouse. Journal of healthcare information management. 2001;15(2):155-164.
8. Houston AL, Chen H, Hubbard SM, Schatz BR, Ng TD, Sewell RR, Tolle KM. Medical Data Mining on the Internet: Research on a Cancer Information System. Artificial Intelligence Review. 1999;13:437-466.
9. *KDnugets* apklausa [interaktyvus]. Prieiga per internetą: <http://www.kdnuggets.com/polls/2011/algorithms-analytics-data-mining.html> [žiūrėta 2014-11-16].
10. Uschold M. Knowledge level modeling: Concepts and terminology. Knowledge Engineering Review. 1998;13(1).
11. Špečkauskienė V, Lukoševičius A. Duomenų tyrybos ir analizės metodų taikymo medicininiams sprendimams rengti metodologija: atvejų tyrimas. Elektronika ir elektrotechnika. 2009;2(90):25-28.
12. Cios KJ, Moore GW. Uniqueness of medical data mining. Artificial Intelligence in Medicine. 2002;26:1-24.
13. Baylis P. Better healthcare with data mining. White paper. Shared Medical Systems Limited, UK; 1999.
14. John Graunt, 1662, Natural and political observations mentioned in a following index, and made upon Bills of Mortality, Citizens of Lonon.
15. Report by the US Congress Office of Technology Assessment. Policy Implications of Medical Information Systems, 1977. Prieiga per internetą: <http://digital.library.unt.edu/ark:/67531/metadc39374/m1/1/> [žiūrėta 2014-11-15].
16. Wasan S, Bhatnagar V, Kaur H. The impact of data mining techniques on medical diagnostics. Data Science Journal. 2006;5:119-126.
17. Standartų sąsajumo specifikacija. HL7/ASTM Implementation Guide for CDA Releae 2 – Continuity of Care Document. HL7 ir ASTM, 2006.
18. Gruninger M, Bodenreider O, Olken F, Obrst L, Yim P. Ontology, taxonomy, folksonomy: Understanding the distinctions. Ontology summit, 2007.
19. Bodenreider O. Ontologies for Mining Biomedical Data. IEEE International Conference on IEEE International Conference on Bioinformatics and Biomedicine; 2008.
20. Tanwani AK, Afridi J, Shafiq MZ, Farooq M. Guidelines to Select Machine Learning Scheme for Classification of Biomedical Datasets. 7th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics; 2009.
21. Standarto aprašymas CRISP-DM 1.0 Step-by-step data mining guide. CRISP-DM 2.0 Special Interest group; 2000. Prieiga per internetą: <http://www.crisp-dm.org/CRISPWP-0800.pdf> [žiūrėta 2013-02-15].

Data Mining in Medicine: applications, challenges and possibilities

Olegas Niakšu
Vilnius University, Institute of Mathematics and Informatics

Summary

Background. Data mining, as a part of knowledge discovery process, expands the limits of statistical analysis and provides news possibilities to researches and practitioners in digital data analysis.

Objectives. The main objective of the study is to review and summarize data mining applications in medicine, with emphases on specific challenges and constraints of the healthcare domain.

Methods. A systematic review of the publications relevant to our research topic was carried out. Qualitative analysis techniques - grounding theory methods, comparison analysis, domain analysis, and taxonomical analysis - were used.

Results and summary. Data mining in healthcare introduces additional issues and challenges to the researchers. Heterogeneous data, patient privacy, big data are the most prominent features. However, the industry standard data mining process models can be applied in medical settings. For the purpose, the data mining application methodology CRISP-DM is described and proposed.

Keywords: data mining, medical information system, healthcare information system.

Correspondence to Olegas Niakšu
Vilnius University
Institute of Mathematics and Informatics
Akademijos str. 4, LT-08663 Vilnius, Lithuania
E-mail: niaksu@acm.org

Received 27 October 2014,
accepted 28 November 2014