Vilniaus universitetas
Duomenų mokslo ir skaitmeninių technologijų institutas
L I E T U V A

INFORMATIKA (09 P)

KLASTERIZAVIMO ALGORITMAI DIDELĖS APIMTIES MEDICINOS DUOMENIMS

Roma Puronaitė

2018 m. spalis

Mokslinė ataskaita DMSTI-DS-09P-2018-1

VU Duomenų mokslo ir skaitmeninių technologijų institutas, Akademijos g. 4,
Vilnius LT-08663
www.mii.lt

Santrauka

Darbe nagrinėjama medicinos duomenų problematika, kuriai spręsti taikytini klasterizavimo metodai. Svarstomos multigrafų, temporalinių (laikinių) grafų ir jungtinio klasterizavimo metodų panaudojimo medicinos duomenims klasterizuoti galimybės.

Reikšminiai žodžiai: elektroninė ligos istorija, elektroniniai sveikatos įrašai, administracinės sveikatos priežiūros paslaugų duomenų bazės, poliligotumas, multigrafas, temporalinis (laikinis) grafas, jungtinis klasterizavimas

Turinys

Contents

1 Įvadas 4
2 Poliligotumo analizė 4
3 Multigrafas ir temporalinis (laikinis) grafas 5
4 Jungtinis klasterizavimas 5
5 Ligoninės informacinės sistemos integracija į nacionalinę elektroninių sveikatos įrašų sistemą 6
6 Išvados 7
7 Literatūra 7

1 Įvadas

Sveikatos priežiūroje elektroninės sveikatos įrašų atsiradimas įnešė daug pokyčių. Be pagrindinės savo paskirties – saugoti duomenis administraciniais tikslais, atsivėrė gana platus antrinio panaudojimo galimybių spektras. [1] Pradedant nuo klasikinių atrankos klinikiniams tyrimams uždavinių iki gilesnės analizės, pacientų ligos istorijos trajektorijų nagrinėjimo. Pastarasis uždavinys yra viena iš galimybių išeiti iš klinikinio tyrimo griežtai apibrėžtos atrankos į „realaus pasaulio“ duomenų analizę. Viena iš tokių plataus susidomėjimo sričių – poliligotumas, pacientų, turinčių kelias lėtines ligas nagrinėjimas [2]. Ligų sąveikos, sąsajos su didesnėmis gydymo išlaidomis, apsunkinta pacientų gyvenimo kokybė, geriausio priežiūros modelio paieška, esamo priežiūros modelio išgryninimas yra susidomėjimo objektas ir sveikatos politikos formuotojams, medicinos praktikams ir mokslininkams. Tuo tarpu papildomas iššūkis ir klausimas yra kaip tokie duomenis turėtų būti analizuojami, kaip apjungiami skirtingi duomenų šaltiniai ir įgalinami bei šiam tikslui kuo labiau išnaudojami elektroniniai sveikatos įrašai, administracinės sveikatos priežiūros duomenų bazės.

Ligoninių informacinių sistemų integravimas į nacionalines sveikatos įrašų sistemas taip pat yra vienas iš iššūkių, kuriems spręsti būtų galima pasitelkti klasterizavimo metodus. [3]

2 Poliligotumo analizė

Poliligotumas – dviejų ir daugiau ligų koegzistavimas pacientui. [2] Būklė imta plačiai nagrinėti dėl aiškių nuorodų į sveikatos priežiūros išlaidų išaugimą poliligotiems pacientams, jų gyvenimo kokybės blogėjimą, poveikį darbingumui ir t.t. Nors poliligotumui ir su juo susijusiems aspektams nagrinėti naudojamos ir tokios priemonės kaip apklausa ar biomedicininiai tyrimai. Vis dėlto daugiausiai žadanti sritis yra elektroniniai sveikatos įrašai ir administracinės sveikatos priežiūros duomenų bazės. Pirmiausiai dėl to, kad jos apima labai didelį būtent su sveikata susijusių parametrų rinkinį ir įtvirtintą poliligotumui vertinti reikalingą žymenį – diagnozes, fiksuotas tam tikro konkretaus sveikatos priežiūros veiksmo (pvz. hospitalizacijos, apsilankymo pas šeimos gydytoją) metu.

Plačiausiai paplitęs analizės būdas – tam tikro periodo pacientų duomenys apibendrinami ir suskaičiuojant kurias ligas pacientas turi, kurių neturi. Toliau nagrinėjamos kombinacijos. Taikomi įvairūs klasterizavimo metodai pacientams arba jų diagnozėms klasterizuoti. [4]

Tačiau kitas aspektas – paciento ligos trajektorijų analizė, nėra dažnas. Akivaizdu, kad trajektorija turėtų būti svarbi. Pacientas nesuserga visomis ligomis iš karto, todėl tų ligų koegzistavimas savaime dar neduoda jokių prielaidų dėl to, kaip būtų galima vertinti priežastingumą. Matome, kad poliligotumas „jaunėja“, poliligotų pacientų proporcija ima augti jau ties 28 metus. [5] Kyla klausimas ar tų pačių ligų rinkinys jaunam pacientui ir vyresniam pacientui yra iš tiesų tapatus. Žingsnis link prevencijos planavimo, pacientui tinkamo gydymo parinkimo, ligos vystymosi suvokimo galėtų būti ligos trajektorijų klasterizavimas.

Be ligos trajektorijos nagrinėjimo apsibrėžiant kažkurį konkretų parametrą (pvz. lėtinių ligos fiksavimas medicinos dokumentuose sveikatos priežiūros veiksmo metu), aktualus ir kitų pacientą galinčių apibrėžti požymių vertinimas. Tai paciento klinikiniai tyrimai, laboratoriniai tyrimai, galbūt ir paciento vaizdai, būklės aprašai. Iš

esmė duomenys, kuriems būdingas nereguliarus išsidėstymas laike ir pasiskirstymas į blokus.

3 Multigrafas ir temporalinis (laikinis) grafas

Nagrinėjant elektroninių sveikatos įrašų ir administracinių sveikatos sistemų panaudojimo poliligotumo analizei galimybes išskirtina, kad labiausiai su šios būklės apibrėžimu susijusi yra diagnozė, kuri fiksuojama kiekvieno sveikatos priežiūros pagrindinio veiksmo metu.

Apsibrėžiant, kad analizės objektas bus paciento trajektorija, galima išskirti tokius požymius:

- *Ligas pacientui galima išdėstyti paminėjimo tvarka ir ši tvarka yra svarbi. (orientuota, žymėta)*
- *Paciento liga gali būti pakartotinai paminėta. (kilpa).*
- *Paciento liga gali būti pakartotinai paminėta kelis kartus.*
- *Paciento ligų sekos dalyje gali būti atsikartojančių fragmentų.*
- *Atstumas tarp ligų yra svarbus.*

Iš to seka, kad paciento ligas galima apsibrėžti kaip žymėtą orientuotą multigrafą.

Kiekvienas pacientas iš dalies dalinasi kažkuria savo multigrafo dalimi su kitu pacientu. Kuo daugiau jų multigrafai sutampa, tuo labiau šie pacientai panašūs.

Klausimas kaip reikėtų įvertinti tai, kad ligos išsidėsto ne tik tam tikra tvarka, bet ir skirtingais laiko tarpais.

Kaip galimas variantas temporalinio (laikinio) grafo idėja [6]. Svorio briaunai parinkimas atliekamas atsižvelgiant į tai, kiek arti ar toli nutolę laiko atžvilgiu yra atitinkamos viršūnės (diagnozės, įvykiai, statusai). Kuo intervalas didesnis, tuo labiau nyksta ryšio svarba.

Temporalinis grafas apibrėžiamas taip:

Laikome, kad turime įvykių sekų aibę $\{s_n : n = 1, \dots, N\}$, O kiekviena seka aprašoma kaip $s_n = ((x_{nl}, t_{nl}): l = 1, \dots, N_L)$, kur L_n yra s_n ilgis, x_{nl}- įvykis, t_{nl}- įvykio laiko momentas sekoje.

Temporalinis grafas G^n iš sekos $s_n$yra orientuotas ir svorinis grafas su įvykių aibe kaip viršūnių aibę {1,..., M}, kur briaunos jungiančios i ir j viršūnes svoris yra apibrėžtas kaip:

$W_{ij}^n = \frac{1}{L_n} \sum_{1 \le p \le q \le L_n} [x_{np} = i \wedge x_{nq} = j] \kappa(t_{nq} - t_{np})$,

kur $\kappa(\cdot)$ yra neaugimo funkcija, kuri apibrėžiama:

$$\kappa(\delta) = \begin{cases} \exp\left(-\frac{\delta}{r}\right) & \delta \le \Delta \\ 0 & \delta > \Delta \end{cases}$$

Parametrai delta ir r parenkami pagal tai, koks laiko intervalas tarp ligų laikytinas prasmingu bei vidutinė trukmė tarp įvykių.

4 Jungtinis klasterizavimas

Kadangi duomenys apie pacientą ateina ne iš vieno šaltinio, duomenų struktūra ir formatas nebūtinai sutampa. Tam reikia tiek specifinio apdorojimo, pvz. transformuoti vienus duomenis į vektorių.

Argumentuojama [7], kad tokie sprendimai gali sąlygoti tikslumo praradimą, kai klasterizuojant kažkuris formatas turi pirmumą.

Siūloma [7] atlikti jungtinį klasterizavimą, kai klasterizuojant atsižvelgiama į jungtinį modelį.

Daromos prielaidos:

- *Abi duomenų dalys yra nepriklausomos, tik turi bendras klasterio žymės.*
- *Modelyje abiejų tipų duomenys laikomi lygiaverčiais ir nelaikomi vienas kito kovariante.*

Laikoma, kad yra *N* objektų, kuriems yra vektoriniai duomenys x_i ir poriniai duomenys y_{ij} (pvz. grafo briaunos jungiančios *i* ir *j* viršūnes svoris), kur *i, j = 1,..., N* indeksai. Tuomet *X* yra *N*x*q* matrica suformuota iš x_i, o *Y* yra *N*x*N* matrica iš y_{ij}. Daroma prielaida, kad *X ir Y turi bendras klasterių žymes C = (c1, ..., cN),* kur *i*-ojo objekto c_i žymė yra klasterio, kuriam priklauso *i*-asis objektas numeris.

Laikoma, kad esant duotajam *C* x_i ir y_{ij} nepriklausomai pasiskirstę pagal savo atskirus skirstinius.

Jungtinė tikėtinumo funkcija yra

$L(X,Y|\Phi,\Psi,C) = \prod_{i=1}^{N} f(x_i|\phi_{c_i}) \cdot \prod_{i=1}^{N}\prod_{j=1}^{N} g(y_{ij}|\psi_{c_i,c_j})$, *kur*

$\Phi = (\phi_1, \dots, \phi_K)$, $\Psi = (\psi_1, \dots, \psi_K)$,speficiniai parametrai, $f(\cdot) ir\ g(\cdot)$ komponentų pasiskirstymo tankio funkcijos.

Toliau daroma prielaida, kad N klasterių žymenų pasiskirstę pagal multinominį skirstinį su tikimybėmis $P = (p_1, \dots, p_k) \in \mathbb{R}^K$, $c_i = k$ su tikimybe $p_k, k = 1, \dots, K$.

Modelis atrodo taip:

$$c_i \sim Multinomis(P),$$
$$x_i|c_i \sim f(x_i|\phi_{c_i},$$
$$y_{ij}|c_i, c_j \sim g\left(y_{ij}\middle|\psi_{c_i,c_j}\right).$$

Pritaikomumas klinikiniam fenotipavimui šiuo atveju būtų atskirų duomenų šaltinių apjungimas. 1 etapo – klasterizavimo pagal ligas (pvz. naudojant temporalinį multigrafą) ir 2 etapo - klasterizavimo pagal kitus požymius (pacientą apibūdinančių savybių vektorius ar matricas).

5 Ligoninės informacinės sistemos integracija į nacionalinę elektroninių sveikatos įrašų sistemą

Šiuo metu Lietuvoje vis dar vyksta ligoninių informacinių sistemų integracija į nacionalinę sveikatos įrašų sistemą. Viena iš didesnių problemų yra ta, kad kuo labiau išvystyta lokali sistema, tuo sudėtingiau suderinti su nacionaline sistema. Tiesioginis panaudojimas sulėtina procesą, tuo pačiu sutrikdo informacinės sistemos veikimą ir taip veikia darbo procesą. Ieškoma sprendimų kaip užtikrinti kuo mažesnį pertraukimą ir stabilų veikimą. [3]

Pagrindiniai iššūkiai [3]:

- *Pranešimų apsikeitimo suderinimas.*
- *Medicinos įrašo perdavimo proceso foniniame režime užtikrinimas.*

Pranešimų apsikeitimo suderinimas sprendžiamas ieškant priemonių, kuriomis lengviausia automatiškai konvertuoti vieno formato žinutes į kito formato.

Medicinos įrašo perdavimo proceso foniniame režime užtikrinimui procesas skaidomas į daug tarpinių statusų, taip užtikrinant, kad nepavykus perdavimui nebus sutrikdytas ligoninės darbuotojų darbas, automatiškai apdorojant galimus scenarijus.

Siekiant dar geriau sureguliuoti procesą, tikslinga taikyti klasterizavimo metodus, pvz. medicinos dokumentų klasterizavimą, proceso įrašų (logų) klasterizavimą.

Šios priemonės galėtų atskleisti galimas silpnas ir stiprias vietas, padėti įvertinti geresnį duomenų perdavimo eiliškumą, blokus ir pan.

6 Išvados

Turėtų būti toliau ieškoma ir bandomi būdai kaip išplėsti metodų pritaikymą apjungiant skirtingų šaltinių, skirtingo tipo duomenys, kaip įvertinti paciento parametrų kitimo laike problema.

7 Literatūra

[1] Yadav, P., Steinbach, M., Kumar, V., & Simon, G. (2018). Mining electronic health records (EHRs): a survey. ACM Computing Surveys (CSUR), 50(6), 85.

[2] Navickas, R., Visockienė, Ž., Puronaitė, R., Rukšėnienė, M., Kasiulevičius, V., & Jurevičienė, E. (2015). Prevalence and structure of multiple chronic conditions in Lithuanian population and the distribution of the associated healthcare resources. European journal of internal medicine, 26(3), 160-168.

[3] Trinkūnas, J., Tuinylienė, E., & Puronaitė, R. (2018, June). Research on Hospital Information Systems Integration to National Electronic Health Record System. In 2018 International Conference BIOMDLORE (pp. 1-6). IEEE.

[4] Prados-Torres, A., Calderón-Larranaga, A., Hancco-Saavedra, J., Poblador-Plou, B., & van den Akker, M. (2014). Multimorbidity patterns: a systematic review. Journal of clinical epidemiology, 67(3), 254-266.

[5] Jurevičienė, E., Onder, G., Visockienė, Ž., Puronaitė, R., Petrikonytė, D., Gargalskaitė, U., ... & Navickas, R. (2018). Does multimorbidity still remain a matter of the elderly: Lithuanian national data analysis. Health Policy.

[6] Liu, C., Wang, F., Hu, J., & Xiong, H. (2015, August). Temporal phenotyping from longitudinal electronic health records: A graph based framework. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 705-714). ACM.

[7] Kong, Y., & Fan, X. (2017). A Bayesian Method for Joint Clustering of Vectorial Data and Network Data. arXiv preprint arXiv:1710.08846.