

2-TUR DIABET KASALLIGINI BASHORATLASH ALGORITMLARINING QIYOSIY TAHLILI

Nasimova Nigoraxon Mizrobovna

Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti

Nasimov Rashid Hamid o'g'li

Toshkent davlat iqtisodiyot universiteti

Annotatsiya

Ushbu maqolada 2-tur diabet kasalligining xavf darajasini baholash, oldindan bashoratlashning sun'iy intellekt algoritmlari tahlil qilingan. 2017-2022 yillar orasida chop etilgan maqolalarni tahliliy o'rganish asosida algoritmlarning bashoratlash aniqligiga ta'sir qiluvchi omillar o'rganilgan va kelgusida qilinishi kerak bo'lgan ishlar ko'rsatilgan.

Аннотация

В этой статье анализируются алгоритмы искусственного интеллекта для оценки уровня риска диабета 2 типа, прогнозирования. На основе аналитического исследования статей, опубликованных в период с 2017 по 2022 год, были изучены факторы, влияющие на точность прогнозирования алгоритмов, и указана работа, которую необходимо проделать в будущем.

Abstract

This article analyzes artificial intelligence algorithms for assessing the risk level of Type 2 diabetes, predictive prediction. Based on an analytical study of articles published between 2017 and 2022, factors affecting predictive accuracy of algorithms have been studied and work to be done in the future has been shown.

Kalit so'zlar

sun'iy intellekt algoritmlari, kasallikni bashoratlash, 2-tur qandli diabet.

Ключевые слова

алгоритмы искусственного интеллекта, прогнозирование заболеваний, сахарный диабет 2 типа.

Keywords

artificial intelligence algorithms, predicting disease, Type 2 diabetes.

Kirish

Jahon sogʻliqni saqlash tashkiloti bergan maʼlumotlarga koʻra bugungi kunda dunyoda 422 milliondan ortiq inson diabet kasalligiga chalingan boʻlib, dunyo miqyosida har yili ushbu kasallik sababli 1,5 million kishi vafot etadi. Diabetga chalinganlar soni 2050 yilga borib 1,5 milliardga yetishi taxmin qilinyapti [1]. Bu insonlarning kam harakatli hayot tarzi, chekish, semirish, notoʻgʻri ovqatlanish kabi zararli odatlar hamda genetik faktorlar, atrof-muhitning ifloshlanishi, aholi oʻrtacha yoshining oshishi kabi boshqa omillar sababli kelib chiquvchi kasallik hisoblanadi. Ushbu kasallikka moyillik genetik va tashqi faktorlar bilan baholanadi. Ushbu kasallikka moyillikni erta aniqlash hamda kasallikka chalinish xavfini baholash orqali xavf faktorlarini kamaytirish va shu orqali kasallikka chalinish xavfini kamaytirish mumkin.

Shu sababli ham oxirgi vaqtlarda diabetning 2-turini oldindan bashorat va xavf darajasini baholash imkonini beruvchi sunʼiy intellekt (SI) algoritmlari ustida koʻplab tadqiqotlar olib borilmoqda. Chunki sunʼiy intellekt algoritmlari katta maʼlumotlar toʻplami bilan ishlay olish natijasida, maʼlumotlar bazasi ichidagi juda murakkab bogʻlanishlarni aniqlashi va shu sababli ham xavf omillarini aniqroq baholash imkoniyatiga egadir.

Ammo shunga qaramay, bugungi kunda taklif etilayotgan sunʼiy intellekt alogirtmlarining ham bir qator kamchiliklari mavju boʻlib, bu maʼlumotlar toʻplamining toʻgʻri shakllantirilmashligi, algoritmning notoʻgʻri tanlanishi modelning aniqligini pasaytiradi yoki real tibbiy holatlarda qoʻllab test qilinmaganligi sababli kamchiliklar yuzaga kelishi mumkin.

Ushbu maqolada 2017-2022 yillar orasida ishlab chiqilgan SI algoritmlari, ularning afzallik va kamchiliklari qiyosiy tahlil qilinadi.

2-tur diabet kasalligini bashorat qiluvchi algoritmlar

Surunkali kasalliklar xavfini bir necha yil oldindan bashoratlash yoki xavflilik darajasini baholash har doim muhim vazifa sanalib kelgan. Shuning uchun bu masalani yechishda turli xil matematik va statistik yondashuvlardan foydalanib kelingan. Masalan, 2-tur diabet xavfini baholash uchun Finnish diabetes risk score (FINDRISC), the German diabetes risk score (GDRS), Korean diabetes risk score (KDS), New Chinese diabetes risk score (NCDRS) va American academy of family physicians (AAFP) risk modellari ishlab chiqilgan. Bu usullarda bemor maʼlumotlaridagi har bir belgi, parametrlar natijaga chiziqli bogʻliq deb hisoblaniladi va ularga alohida qiymat berib, qiymatlar yigʻindisi shunchaki qoʻshiladi [1]. Ammo bunday holatda risk omillari va risk xavfi orasidagi noxiziq va murakkab bogʻliqlik boʻlishi mumkin, bundan tashqari belgilar orasida ham oʻzaro bogʻliqliklar ham boʻlishi mumkin. Ammo, yuqoridagi usullarda haddan tashqari soddalashtirishlar qilinganligi sababli, baʼzi

muhim ma'lumotlarning yo'qolishi kuzatilgan. Shu sababli ham so'nggi yillarda bu sohada sun'iy intellekt algoritmlaridan foydalanish ustida ko'plab tadqiqotlar olib borilmoqda. Natijada, qator algoritmlar ishlab chiqildi: Jumladan, Tasodifiy o'rmon klassifikatori (Random Forest classifier - RF), Gradiyentni kuchaytiruvchi qarorlar daraxti (Gradient Boosting decision tree - GBDT), Logistik regressiya (Logistic regression - LR), chuqur neyron tarmoqlari (Deep neural network - DNN), K-eng yaqin qo'shnilar (K-nearest neighbors - kNN), Sodda Bayes daraxti (Naive Bayes tree), qo'llovchi vektor mashinasi (Support vector machine - SVM) va ularning turli kombinatsiyalaridan foydalanilgan, 1- jadvalda qilingan tadqiqot ishlarining qisqacha ma'lumotlari keltirilgan.

Tadqiqot metodologiyasi

Ushbu tahliliy maqolani yozish uchun Web of Science, PubMed bazalaridagi tahliliy maqolalar ko'rib chiqildi. Bunda "Predictive model", "Artificial Intelligent predictive algorithm", "Machine learning predictive algorithm", "Deep learning methods to predict risc" va "Predicting type 2-diabetes" kalit so'zlaridan foydalanildi. Maqolalar avval faqat nomiga va tahliliy maqola ekanligiga qarab ajratib chiqilgan. Ular ichida tahlil qilingan maqolalar ham alohida tartiblandi. Kasallik xavfini oldindan bashoratlash emas, kasallikni erta aniqlash, kasallikning yashirin belgilarini aniqlash vazifasi bajarilgan maqolalar chiqarib tashlandi.

Algoritmlar tahlili

Bugungi kungacha taklif etilgan algoritmlarni o'zaro taqqoslash bir qancha murakkabliklarga ega. Bunga bir qancha omillar sabab bo'ladi:

Algoritmlarni baholash uchun turli xil baholash mezonlari ishlatish. Ba'zi algoritmlar [3,9,10] faqat aniqlik darajasi bilan va ba'zilar faqat AUC qiymati bilan baholangan [7,8,12,13], 1-jadval. Ba'zilar [5,6] esa butunlay boshqa qiymat bilan baholangan.

Turli ma'lumotlar to'plamidan foydalanish. 2-tur diabetni oldindan xavf darajasini baholash uchun hali kasallikka chalinmagan sog'lom insonlarni bir necha yil davomida ulardagi o'zgarishlarni monitoring qilish va ularning ma'lumotlar to'plamini ishlab chiqish zarur. Bu juda murakkab jarayon bo'lib, shu sababli ham bunday ma'lumotlar to'plamlari hozirgi kunda juda kam. Mavjudlari esa bemorlarning xavfsizligini ta'minlash maqsadida foydalanish uchun taqdim etilmaydi.

Shuning uchun ham tadqiqotchilar o'z algoritmlarini faqat o'zlari to'plagan ma'lumotlar to'plamidagina sinab ko'rishlari mumkin, boshqa ma'lumotlar to'plamida test qilish orqali turli algoritmlarning imkoniyatlarini taqqoslash imkoniyati cheklangan.

Mavjud ma'lumotlar to'plami bir biri bilan hajmi ko'rsatgichlariga ko'ra farq qiladi: ba'zilar [1,4,9] yetarlicha katta bo'lsada, ba'zilar juda kichik [2,3,6] va shu sababli bu ma'lumotlar to'plamida muvozanatlashmagan to'plamida o'qitilgan va test qilingan algoritmning keltirilgan aniqligi shubhali hisoblanadi. Ko'pgina ma'lumotlar to'plamida [1,4,5,7,8] juda katta muvozanatlashmagan mavjud bo'lib, algoritmni o'qitishdan oldin 2 ta sinfga tegishli ma'lumotlarni muvozanatlashmagan usulidan faqat [4,5,8] ishlardagina foydalanilgan. Yana bir muhim sabab ma'lumotlar to'plamidagi bemorlar yoshi, jinsi, etnik mansubligi turlichaligida. Tadqiqotchilar tadqiqot sub'yekti sifatida turli xil yoshdagi guruhlardan ma'lumotlar to'plashgan. Ma'lumki, 2-tur

diabetga chalinishi uchun yosh ham xavf omillaridan biri hisoblanadi. 18-35 yoshlilar uchun katta tana vazni xavf omili bo‘lmasligi, 60-75 yoshlilar uchun esa juda katta xavf omili hisoblanishi mumkin. Shu sababli ham turli ma’lumotlar to‘plamida o‘qitilgan algoritmlar biror xavf omilini turlicha baholashi mumkin.

1-jadval

Taklif etilgan algoritmlar va ularda foydalanilgan ma’lumotlar to‘plami haqida ma’lumotlar

№	Muallif / (yil)	Algoritm nomi	Ma’lumotlar to‘plami hajmi (Jami/Diabet)	Baholash	Kuzatish vaqti	Belgi xususiyatlar
1	[1] / (2017)	NB, RF, Logistik model daraxti, J48 qarorlar daraxti	32,555/5099	Aniqlik: 0.83, 0.84, 0.79 AUC(ROC): 0.67, 0.66, 0.69, 0.63	5 yil	13 ta
2	[2] / (2020)	Ko‘p misolli o‘rganishni kuchaytirish	256/256	Aniqlik: 0.83 AUC (ROC): 0.89	9 yil	49
3	[3] / (2017)	K2 strukturali algoritmi	748/374	Aniqlik: 0.82	2 yil	16
4	[4] / (2019)	DNN	9948/1904	Aniqlik: 0.84 AUC (ROC): 0.84	4 yil	365
5	[5] / (2019)	Cox proporsional xavf regressiya modeli	6144/331	Xavf nisbati AUC: 0.75	3 yil	5
6	[6] / (2017)	Ikkilik xavflarni tasniflovchi eng kichik kvadrat regressiyasi	543/146	Imkoniyat nisbati aniqligi (AUC): 0.77	10 yil	568
7	[7] / (2019)	J48 qarorlar daraxti, NB	4403/358	AUC (ROC): 0.79	10 yil	5
8	[8] / (2019)	LR, chiziqli diskriminant tahlili, kvadratik diskriminant tahlili, k-NN	8454/404	AUC (ROC): 0.78, 0.77, 0.76, 0.77	5 yil	28
9	[9] / (2019)	k-NN, SVM, DT, RF, GBDT, NN, NB	10,000	Aniqlik: 0.81, 0.81, 0.79, 0.82, 0.82, 0.82, 0.66	3 yil	6
10	[10] / (2019)	Gauss radial asosli SVM	1438/161	Aniqlik: 0.96	7,5	10
11	[11] / (2020)	k-NN	1647/214	Aniqlik: 0.977 AUC (ROC): 0.89	2 yil	3
12	[12] / (2019)	LR, k-NN, SVM	1837/647	AUC (ROC): 3-yil: 0.74, 0.83, 0.73 5-yil: 0.72, 0.82, 0.68 7-yil: 0.70, 0.79, 0.71	3,5,7 yil	8
13	[13] / (2021)	RF, LR, Adaboost, k-NN, DT	2278/833	AUC (ROC): 0.91, 0.91, 0.90, 0.86, 0.78	2 oy	19

Risk faktor sifatida turli xil belgi xususiyatlarni tanlab olish. Algoritmning aniqligi unga taqdim etilgan kirish qiymatlari soni va sifatiga kuchli bog‘liqdir. Ammo 1-jadvaldan ko‘rinib turibdiki, tadqiqotchilar kirish qiymati sifatida 3 tadan [11] 568 tagacha [6] belgi xususiyatlardan foydalangan. Bunday holatlarda algoritm aniqligini taqqoslash imkoniyati kamayishiga olib kelishi mumkin.

Xavf darajasini baholashning yilga bog‘liqligi. Kasallikning xavf darajasini baholash kasallikka chalinishdan necha yil oldin ekanligiga bog‘liq bo‘ladi [12]. Ammo, olib borilgan tadqiqotlar turli xil yillar orasida olib borilganligi sababli ularning aniqligini to‘g‘ridan to‘g‘ri taqqoslash mumkin emas. Shu sababli ushbu ishda bir xil ma’lumotlar to‘plami o‘qitilgan va test qilingan algoritmlarning qiyosiy tahlili bilan cheklaniladi. Bir nechta algoritmlar ayni bitta ma’lumotlar to‘plamida test qilinganligini ko‘rishimiz mumkin [1,8,9,12,13]. 1-advaldan ko‘rinib turibdiki, turli sinflarga tegishli ma’lumotlar muvozanatlashmagan ma’lumot to‘plamida [1,13] o‘qitilgan LR, RF algoritmlari k-NN algoritmgiga nisbatan biroz yuqori natijaga erishgan. Ammo muvozanatlashgan ma’lumotlar to‘plamida [12] k-NN algoritmi LR va RF algoritmlariga qaraganda sezilarli darajada yuqori natijani ko‘rsatgan.

k-NN algoritmi bashoratlash bo‘yicha eng yuqori natijani ko‘rsatgan bo‘lsada, bu ishda faqat diabetga moyillar (prediabetes)dan iborat insonlarning diabetga chalinish

xavfi o'rganilgan [11]. Bunday kuzatiluvchilardagi xavf omillar sog'lom insonlardagiga qaraganda kuchli ta'sir etadi. Undan tashqari bu ishda boshqa ishlardan farqli ravishda 3 ta parametrning uzluksiz o'zgarishi haqidagi davriy ma'lumotlardan foydalanilgan. Undan tashqari ushbu ma'lumotlar to'plami nisbatan kichik bo'lgani sababli, ko'rsatilgan aniqlikka asosan k-NN algoritmi eng yuqori aniqlikni ko'rsatgan deya olmaymiz.

jadvaldan yana shunday xulosa chiqarish mumkinki, kasallikni bashoratlash aniqligi necha yil oldindan amalga oshirilgan bashoratlashga ham bog'liq ekan. Kichik vaqt oralig'ida bashoratlash usullarida ayni deyarli barcha algoritmlar yuqori natijaga erishganini ko'rishimiz mumkin. 5 yil oldindan bashoratlashda 80% dan yuqori natijaga erishga RF, LR, k-NN [1] algoritmlari, 2 oy oldindan bashoratlash [13] da 90 % dan yuqori natijalarni ko'rsatgan.

Shuningdek, bashoratlash aniqligi belgi xususiyatlarning soniga emas, tanlangan algoritmgaga bog'liqligini ko'rishimiz mumkin. Bir xil vaqt oldindan bashoratlovchi va hajmi bo'yicha uncha katta farq qilmaydigan ma'lumotlar to'plamida o'qitilgan algoritmlarning [6] birida juda kam (5 ta), boshqasida [5] juda ko'p (568 ta) belgi xususiyatdan foydalanilgan bo'lsada, J48 qarorlar daraxti algoritmi yuqoriroq natijani ko'rsatgan.

Xulosa

Ushbu maqolada ko'rib chiqilgan, 2-tur diabetni bashoratlash algoritmlarni tahlil qilib, xulosa qilindiki, algoritmlarning aniqligi ma'lumotlar to'plamining muvozanatlashganligiga, necha yil oldin bashoratlashga bog'liq ekan. Algoritm aniqligi ma'lumotlar to'plamidagi belgi xususiyatlar soniga emas, belgi-xususiyatlarning muhimlik darajasiga ko'proq bog'liq. Bir xil ma'lumotlar to'plamida o'qitilgan algoritmlar orasida RF eng yuqori natijani ko'rsatishi mumkin.

Ma'lumotlar to'plamining xavfsiz nuqtai nazaridan tarqalib ketmasligi uchun yuqorida tahlil qilingan algoritmlarni birortasini boshqa tadqiqotchilar uchun test qilib tekshirib ko'rish imkoniyati mavjud emas. Buning uchun ayni o'sha belgi-xususiyatlarga ega ma'lumotlar to'plami talab etiladi. Shu sababli ham butun dunyo tadqiqotchilari foydalanishi mumkin bo'lgan, bemorlar shaxsiy daxlsizligiga daxl qilmaydigan sintetik ma'lumotlar to'plamini yaratish dolzarb vazifa bo'lib qolmoqda.

Foydalanilgan adabiyotlar ro'yxati

1. Alghamdi M, Al-Mallah M, Keteyian S, Brawner C, Ehrman J, Sakr S. Predicting diabetes mellitus using smote and ensemble machine learning approach: the henry ford exercise testing (ft) project. PLoS ONE. 2017;12(7):e0179805.
2. Bernardini M, Morettini M, Romeo L, Frontoni E, Burattini L. Early temporal prediction of type 2 diabetes risk condition from a general practitioner electronic health record: a multiple instance boosting approach. Artif Intell Med. 2020;105:101847.
3. Xie J, Liu Y, Zeng X, Zhang W, Mei Z. A Bayesian network model for predicting type 2 diabetes risk based on electronic health records. Modern Phys Lett B. 2017;31(19–21):1740055.
4. Nguyen BP, Pham HN, Tran H, Nghiem N, Nguyen QH, Do TT, Tran CT, Simpson CR. Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. Comput Methods Programs Biomed. 2019;182:105055.

5. Arellano-Campos O, Gómez-Velasco DV, Bello-Chavolla OY, Cruz-Bautista I, Melgarejo-Hernandez MA, et al. Development and validation of a predictive model for incident type 2 diabetes in middle aged Mexican adults: the metabolic syndrome cohort. *BMC Endocr Disord*. 2019;19(1):1–10.
6. Peddinti G, Cobb J, Yengo L, Froguel P, Kravić J, Balkau B, Tuomi T, Aittokallio T, Groop L. Early metabolic markers identify potential targets for the prevention of type 2 diabetes. *Diabetologia*. 2017;60(9):1740–50.
7. Perveen S, Shahbaz M, Keshavjee K, Guergachi A. Metabolic syndrome and development of diabetes mellitus: predictive modeling based on machine learning techniques. *IEEE Access*. 2019;7:1365–75.
8. Choi BG, Rha S-W, Kim SW, Kang JH, Park JY, Noh Y-K. Machine learning for the prediction of new-onset diabetes mellitus during 5-year follow up in non-diabetic patients with cardiovascular risks. *Yonsei Med J*. 2019;60(2):191.
9. Akula R, Nguyen N, Garibay I. Supervised machine learning based ensemble model for accurate prediction of type 2 diabetes. In: 2019 SoutheastCon. 2019.
10. Abbas H, Alic L, Erraguntla M, Ji J, Abdul-Ghani M, Abbasi Q, Qaraqe M. Predicting long-term type 2 diabetes with support vector machine using oral glucose tolerance test. *bioRxiv*.2019.
11. Garcia-Carretero R, Vigil-Medina L, Mora-Jimenez I, et al. Use of a k-nearest neighbors model to predict the development of type 2 diabetes within 2 years in an obese, hypertensive population. *Med Biol Eng Comput*. 2020;58(5):991–1002.
12. Farran B, AlWotayan R, Alkandari H, Al-Abdulrazzaq D, Channanath A, Thanaraj TA. Use of non-invasive parameters and machine-learning algorithms for predicting future risk of type 2 diabetes: a retrospective cohort study of health data from Kuwait. *Front Endocrinol*. 2019;10:624.
13. Boutilier JJ, Chan TCY, Ranjan M, Deo S. Risk stratification for early detection of diabetes and hypertension in resource-limited settings: machine learning analysis. *J Med Internet Res*. 2021;23(1):20123.