TIBBIY TASHXIS QO‘YISH MASALALARINI YECHISHNING NAZARIY ASOSLARI

Eshboyev Erkin Abdirashidovich[1], Klicheva Firuza Gulmuratovna[2], Daminova Shaxzoda Egamberdi qizi[3], Abdirashidov Shohruhmirzo Erkin o’g’li[4]

[1]dotsent, Qarshi DU,
[2]t.f.f.d., v.b. dotsent, Qarshi DU,
[3]magistr Qarshi DU,
[4]O’zMU
E-mail: vipklicheva@mail.ru

KEYWORDS

tibbiy tashxis, sun’iy intellekt, kichik hajmli tanlanma, ko‘p sinfli tasniflash, aralash turdagi alomatlar, Gauer o‘xshashlik o‘lchovi, regulyarizatsiya, diskriminant tahlil

ABSTRACT

Maqolada kichik hajmli tibbiy tanlanmalar asosida tashxis qo‘yish masalalarini yechishning nazariy asoslari ko’rib o’tilgan. Ya’ni tibbiy tashxis masalalari sun’iy intellekt nuqtai nazaridan ma’lumot modalligi, bilim manbai, tanlanma hajmi va qaror qabul qilish tartibi bo‘yicha tasniflangan. Tashxis masalasining formal qo‘yilishi xatoliklar qiymati asimmetrik bo‘lgan klassifikator qabul qiladigan qarorlar natijasida kutiladigan o‘rtacha yo‘qotish (bayescha xavf) va Yu.I.Juravlyovning algebraik yondashuvi doirasida berilgan, kichik hajmli tanlanma holati esa obyektlar soni, alomatlar soni va sinflar soni orasidagi munosabatlar orqali aniqlangan. Ko‘p sinfli tashxis masalalarining o‘ziga xosligi, ya’ni sinflar nomutanosibligi, ierarxik tuzilma, “bittaga qarshi qolganlar”, “bittaga qarshi bitta” va dekompozitsiyalar hamda baholash metrikalari tahlil qilingan. Aralash turdagi alomatlar muammosi o‘xshashlik o‘lchovlari nuqtai nazaridan ko‘rib chiqilgan. Ikkita hisoblash tajribasida klassik chiziqli diskriminant tahlili tanlanma hajmi kichraygan sari keskin yomonlashishi (har sinfda 5 ta obyektda makro-F1 = 0,744), Ledoit–Wolf qisqartirishi esa uni 0,934 gacha tiklashi, aralash turdagi ma’lumotlarda alomatlar turini hisobga olmaydigan Evklid masofasi makro-F1 qiymatini 0,362 gacha tushirib yuborishi ko‘rsatib o’tilgan. Olingan natijalar kichik hajmli tibbiy ma’lumotlar uchun regulyarizatsiyalangan va alomatlar turini hisobga oluvchi usullarni tanlash bo‘yicha nazariy asos bo‘lib xizmat qiladi.

Kirish

Tibbiy tashxis – bu bemor haqidagi alomatlar (simptomlar, laborator va instrumental ko‘rsatkichlar, anamnez ma’lumotlari) majmuasi asosida uning holatini oldindan belgilangan nozologik sinflardan biriga kiritish jarayonidir. Bu jarayonni matematik modellashtirishga dastlabki urinishlar o‘tgan asrning 50-yillari oxiriga borib taqaladi. O’sha yillarda R.Ledli va L.Lastedning ishida tashxis mantiqiy va ehtimoliy xulosa chiqarish masalasi sifatida qo‘yilgan edi [1]. Keyinchalik MYCIN kabi bilimlarga asoslangan ekspert tizimlari [2], so‘ngra esa ma’lumotlardan o‘rganuvchi mashinali o‘qitish usullari keng tarqaldi [3–5].

So‘nggi o‘n yillikda chuqur neyron tarmoqlar katta hajmli tasvir to‘plamlarida mutaxassis darajasidagi aniqlikka erishdi [6–8]. Biroq bunday natijalar yuz minglab belgilangan namunalarga tayanadi. Klinik amaliyotda esa ko‘pincha teskari holat kuzatiladi ya’ni kam uchraydigan kasalliklar, yangi tashxis usullari, bitta klinika ma’lumotlari yoki qimmat laborator tahlillar bo‘yicha tanlanma hajmi bir necha o‘nlab yoki yuzlab bemordan oshmaydi, alomatlar soni esa o‘nlab, ba’zan yuzlab bo‘ladi. Bunday sharoitda statistik baholashlar beqarorlashadi, modellar o‘qitish tanlanmasiga haddan tashqari moslashadi va sifatni ishonchli baholashning o‘zi alohida muammoga aylanadi [9-11].

Masalani murakkablashtiruvchi yana ikki omil mavjud. Birinchidan, real tashxis masalalari aksariyat hollarda ko‘p sinfli bo‘lib, sinflar hajmlari keskin farq qiladi va ular orasida ierarxik munosabatlar mavjud. Ikkinchidan, tibbiy alomatlar tabiatan turli tipli bo‘lib, arterial bosim miqdoriy, kasallik bosqichi tartibli, chekish odati

(cc) BY

binar, qon guruhi esa nominal alomatdir. Ko‘pgina klassik usullar faqat bir tipli (odatda miqdoriy) alomatlar uchun ishlab chiqilgan.

Ushbu maqolaning maqsadi kam hajmli tibbiy tanlanmalarda tashxis qo‘yish masalalarini yechishning nazariy asoslarini tizimlashtirishdan iborat. Buning uchun quyidagi vazifalarni bajarish kerak bo’ladi:

- tibbiy tashxis masalalarini sun’iy intellekt nuqtai nazaridan tasniflash;
- tashxis masalasining formal qo‘yilishini berish va kam hajmli tanlanma holatini aniqlashtirish;
- ko‘p sinfli tashxis masalalarining o‘ziga xosliklarini ochib berish;
- aralash tipli alomatlar muammosini tahlil qilish;
- asosiy nazariy xulosalarni hisoblash tajribalari bilan tasdiqlash.

Materiallar va usullar

Ish nazariy-tahliliy xususiyatga ega bo‘lib, unda tasvirlarni tanish nazariyasi [12, 13], statistik o‘qitish nazariyasi [3, 4], ko‘p sinfli tasniflash nazariyasi [14–16] va heterogen ma’lumotlar uchun o‘xshashlik o‘lchovlari nazariyasi [17] bo‘yicha asosiy manbalar umumlashtirildi. Nazariy xulosalarni tekshirish uchun ikkita hisoblash tajribasi o‘tkazildi.

2.1. Tibbiy tashxis masalasining formal qo‘yilishi

Bemorlar (obyektlar) to‘plami M berilgan bo‘lsin. Har bir obyekt n ta alomat bilan tavsiflanadi:

$$x = (x_1, x_2, \dots, x_n) \in X = X_1 \times X_2 \times \cdots \times X_n, \quad (1)$$

bu yerda X_j - j-alomatning qiymatlar to‘plami. Aralash tipli ma’lumotlarda X_j turli tabiatga ega bo‘ladi: $X_j \subseteq \mathbb{R}$ (miqdoriy), $X_j = \{1,2,\dots,r_j\}$ tartib munosabati bilan (tartibli), $X_j = \{0,1\}$ (binar) yoki tartibsiz chekli to‘plam (nominal). Bundan tashqari, alomat qiymati noma’lum bo‘lishi mumkin, shuning uchun buni $x_j = \varnothing$ deb belgilaymiz.

M to‘plam o‘zaro kesishmaydigan l ta sinfga (tashxislarga) bo‘lingan:

$$M = \bigcup_{k=1}^{l} K_k, \qquad K_i \cap K_j = \varnothing \ (i \neq j). \quad (2)$$

O‘qitish tanlanmasi

$$S_m = \{(x^{(i)}, y^{(i)})\}_{i=1}^{m}, \qquad y^{(i)} \in Y = \{1, \dots, l\}, \quad (3)$$

ko‘rinishida berilgan, bunda m_k - k-sinfga tegishli obyektlar soni, $\sum_{k=1}^{l} m_k = m$. Tashxis masalasi - $A: X \to Y$ algoritmini (qaror qoidasini) qurishdan iborat bo‘lib, u yangi bemor uchun to‘g‘ri tashxisni imkon qadar kichik xavf bilan qo‘yishi kerak.

Masalaning Bayescha qo‘yilishi. (x, y) juftliklar noma’lum $P(x, y)$ taqsimotdan olinadi deb faraz qilinadi. Tibbiyotda xatoliklar qiymati asimmetrikdir ya’ni xavfli kasallikni o‘tkazib yuborish soxta signaldan ancha qimmatga tushadi. Shuning uchun $\Lambda = (\lambda_{kj})_{l \times l}$ yo‘qotishlar matritsasi kiritiladi, bunda λ_{kj} - haqiqiy sinf k bo‘lganda j tashxisini qo‘yish qiymati. Algoritm xavfi

$$R(A) = \sum_{k=1}^{l} \sum_{j=1}^{l} \lambda_{kj} \ P(A(x) = j, \ y = k) \quad (4)$$

ga teng, uni minimallashtiruvchi bayescha optimal qoida esa

$$A^*(x) = \underset{j \in Y}{argmin} \sum_{k=1}^{l} \lambda_{kj} \ P(K_k \mid x) \quad (5)$$

ko‘rinishga ega. $\lambda_{kj} = 1 - \delta_{kj}$ bo‘lganda (5) maksimal aposterior ehtimollik qoidasiga aylanadi. Klinik amaliyotda “tashxis qo‘yishdan voz kechish” (bemorni qo‘shimcha tekshiruvga yo‘naltirish) imkoniyati muhim. Buni $l + 1$-qaror sifatida $\lambda_{k,0} = \lambda_0$ qiymat bilan kiritish mumkin. U holda qaror $\max_k P(K_k \mid x) < 1 - \lambda_0$ bo‘lganda rad etiladi.

Masalaning algebraik qo‘yilishi. Yu. I. Juravlyov yondashuvida [12, 13] algoritm har bir obyekt uchun axborot vektorini hosil qiladi:

$$A(x) = (\alpha_1(x), \dots, \alpha_l(x)), \qquad \alpha_k(x) \in \{0,1,\Delta\}, \quad (6)$$

bunda $\alpha_k(x) = 1$ - $x \in K_k$, $\alpha_k(x) = 0$ - $x \notin K_k$, Δ - algoritm qaror qabul qilishdan voz kechdi.

(cc) BY

Algoritm $A = C \circ B$ ko‘rinishda bo’ baholovchi operator B (sinflarga tegishlilik baholari $\Gamma_k(x)$ ni hisoblaydi) va hal qiluvchi qoida C kompozitsiyasi sifatida ifodalanadi. Bu yondashuvning tibbiyot uchun afzalligi shundaki, Δ javobi klinik «noaniq holat» bilan bevosita mos keladi, baholar asosidagi algoritmlar (masalan, hisoblash baholari algoritmlari) esa kichik tanlanmalarda ham qo‘llanilishi mumkin.

Empirik xavfni minimallashtirish. $P(x, y)$ noma’lum bo‘lgani uchun (4) o‘rniga regulyarizatsiyalangan empirik xavf minimallashtiriladi:

$$\hat{A} = \underset{A \in \mathcal{A}}{argmin} \left[\frac{1}{m} \sum_{i=1}^{m} L\left(y^{(i)}, A\left(x^{(i)}\right)\right) + \gamma\, \Omega(A) \right], \qquad (7)$$

bu yerda $\mathcal{A}$ - algoritmlar modeli, L - yo‘qotish funksiyasi, Ω - murakkablik jarimasi, $\gamma \geq 0$ - regulyarizatsiya koeffitsiyenti.

2.2. Kam hajmli tanlanma tushunchasi

V. N. Vapnik nazariyasiga ko‘ra [3], VC-o‘lchami h bo‘lgan algoritmlar sinfi uchun $1 - \eta$ ehtimollik bilan quyidagi baho o‘rinli:

$$R(A) \leq R_{emp}(A) + \sqrt{\frac{h\left(ln\frac{2m}{h}+1\right) - ln\frac{\eta}{4}}{m}}. \qquad (8)$$

(8) dan ko‘rinadiki, umumlashtirish xatosining yuqori chegarasi h/m nisbati bilan boshqariladi. n o‘lchovli fazodagi chiziqli tasniflagichlar uchun $h = n + 1$, demak m alomatlar soniga yaqin yoki undan kichik bo‘lsa, empirik xato haqiqiy xato haqida deyarli hech narsa demaydi. Shu munosabat bilan quyidagi ta’rifni qabul qilamiz. Agar quyidagi shartlardan kamida bittasi bajarilsa, S_m tanlanma kam hajmli deyiladi:

$$\frac{m}{n} < c_1, \qquad \min_k m_k < c_2, \qquad \min_k m_k < n, \qquad (9)$$

bunda c_1 va c_2 - model murakkabligiga bog‘liq chegaralar. Amaliy tavsiyalarda chiziqli modellar uchun $m/n \geq 5 \div 10$ talab qilinadi [9], klinik bashorat modellarida esa har bir parametrga to‘g‘ri keladigan hodisalar soni (EPV) hisobga olinadi. (9) dagi uchinchi shart alohida ahamiyatga ega: u bajarilganda sinf ichidagi kovariatsiya matritsasining bahosi xos (singulyar) bo‘lib qoladi.

Haqiqatan ham, chiziqli diskriminant tahlilida (LDA) umumiy kovariatsiya matritsasi

$$\hat{\Sigma} = \frac{1}{m-l} \sum_{k=1}^{l} \sum_{y^{(i)}=k} \left(x^{(i)} - \hat{\mu}_k\right)\left(x^{(i)} - \hat{\mu}_k\right)^T \qquad (10)$$

ko‘rinishda baholanadi va uning rangi $\operatorname{rank}\hat{\Sigma} \leq m - l$. Demak, $m - l < n$ bo‘lganda $\hat{\Sigma}^{-1}$ mavjud emas, $m - l$ qiymati n ga yaqin bo‘lganda esa matritsa yomon shartlangan bo‘ladi. Bu holat G. Xyuz aniqlagan “cho‘qqi hodisasi” (peaking phenomenon) bilan bog‘liq ya’ni tanlanma hajmi o‘zgarmas bo‘lganda alomatlar sonini oshirish ma’lum chegaradan so‘ng aniqlikni pasaytiradi. Kovariatsiya bahosini barqarorlashtirishning standart usuli bu qisqartirish (shrinkage) hisoblanadi [8, 9]:

$$\hat{\Sigma}(\rho) = (1 - \rho)\hat{\Sigma} + \rho\, \nu I, \qquad \nu = \frac{tr\hat{\Sigma}}{n}, \quad \rho \in [0,1], \qquad (11)$$

bunda ρ koeffitsiyenti Ledoit–Wolf formulasi bo‘yicha analitik tanlanadi [18]. Shunga o‘xshash maqsadga L_1 (LASSO) va L_2 regulyarizatsiyalari bilan ham erishiladi [19].

2.3. Aralash toifali alomatlar uchun o‘xshashlik o‘lchovi

Turli tipli alomatlarni yagona fazoda taqqoslash uchun Gauer o‘xshashlik koeffitsiyentidan [17] foydalaniladi. Ikki obyekt orasidagi masofa

$$d_G(x, z) = \frac{\sum_{j=1}^{n} w_j\, \delta_j(x,z)\, d_j(x_j, z_j)}{\sum_{j=1}^{n} w_j\, \delta_j(x,z)} \qquad (12)$$

ko‘rinishda aniqlanadi, bunda w_j - alomat vazni, $\delta_j(x, z) = 0$, agar x_j yoki z_j noma’lum bo‘lsa (yoki asimmetrik binar alomatda ikkalasi ham 0 bo‘lsa), aks holda $\delta_j(x, z) = 1$. Alomat bo‘yicha qismiy masofa uning tipiga qarab tanlanadi:

(cc) BY

$$d_j(x_j, z_j) = \begin{cases} \frac{|x_j - z_j|}{R_j}, & \textit{miqdoriy alomat,} \\ \frac{|r(x_j) - r(z_j)|}{r_j - 1}, & \textit{tartibli alomat,} \\ [x_j \neq z_j], & \textit{binar va nominal alomat,} \end{cases} \quad (13)$$

bu yerda $R_j = \max_i x_j^{(i)} - \min_i x_j^{(i)}$ - j-alomatning tanlanmadagi qiymatlar diapazoni, $r(\cdot)$ - tartib darajasi, $[\cdot]$ - Ayverson qavsi. (12) dagi δ_j ko‘paytuvchi tushirib qoldirilgan qiymatlarni to‘ldirmasdan ishlash imkonini beradi, bu esa kichik tanlanmalarda ayniqsa muhim, chunki to‘ldirish (imputatsiya) modelini ishonchli baholash uchun ham ma’lumot yetarli bo‘lmaydi [20].

2.4. Tibbiy tashxis masalalarining sun’iy intellekt nuqtai nazaridan tasnifi

Adabiyotlar tahlili asosida tibbiy tashxis masalalarini besh mezon bo‘yicha tasniflash taklif etiladi (1-jadval). Bu mezonlar o‘zaro mustaqil bo‘lib, har bir aniq masala ularning har biri bo‘yicha alohida o‘rin egallaydi. Masalan, kam uchraydigan irsiy kasallikni laborator ko‘rsatkichlar asosida aniqlash ko‘p sinfli, jadval ko‘rinishidagi, ma’lumotlarga asoslangan, kam hajmli va rad etish imkoniyatli masaladir.

1-jadval.

Tibbiy tashxis masalalarining sun’iy intellekt nuqtai nazaridan tasnifi

T/R	Mezon	Masala turi	Tibbiy misol	Asosiy usullar
1	Chiqish tuzilishi	Binar	Kasallik bor/yo‘q (skrining)	Logistik regressiya, SVM
		Ko‘p sinfli	Anemiya turini aniqlash	LDA, OvR/OvO, ECOC, softmax [12–14]
		Ko‘p belgili (multi-label)	Yondosh kasalliklar	Binar dekompozitsiya, belgilar zanjiri [34]
		Tartibli	Kasallik bosqichi	Tartibli regressiya
		Ierarxik	XKT-10 bo‘yicha tasniflash	Daraxtsimon tasniflagichlar
2	Ma’lumot modalligi	Jadval	Laborator va klinik ko‘rsatkichlar	Daraxtlar ansambllari, chiziqli modellar [26]
		Tasvir, signal, matn	Rentgen, EKG, epikriz	Chuqur neyron tarmoqlar [21, 23]
3	Bilim manbai	Bilimga asoslangan	Ekspert qoidalari	Ekspert tizimlari [25]
		Ma’lumotlarga asoslangan	Tarixiy kasallik tarixlari	Mashinali o‘qitish [3–5]
		Gibrid	Qoidalar + o‘qitish	Bayes tarmoqlari, cheklangan modellar
4	Tanlanma hajmi	Katta ($m \gg n$)	Populyatsion registrlar	Murakkab modellar, chuqur o‘qitish
		Kam hajmli ((9) shart)	Kam uchraydigan kasalliklar	Regulyarizatsiya, baholar algoritmlari [1, 8–10]
5	Qaror rejimi	Majburiy qaror	Avtomatik triaj	(5) qoida
		Rad etish imkoniyatli	Qo‘shimcha tekshiruvga yo‘naltirish	(6) dagi Δ javob

(cc) BY

Ko‘p sinfli tashxis masalalarining o‘ziga xosligi. Tahlil ko‘p sinfli tibbiy tashxis masalalarining kam hajmli tanlanmalarda keskinlashadigan beshta o‘ziga xosligini ko‘rsatdi.

Birinchidan, sinflar sonining ortishi bilan har bir sinfga to‘g‘ri keladigan ma’lumot kamayadi. Agar umumiy hajm m o‘zgarmas bo‘lsa, o‘rtacha $m_k = m/l$ bo‘ladi va (9) dagi $\min_k m_k < n$ sharti binar holga qaraganda tezroq bajariladi. Kvadratik diskriminant tahlilida har bir sinf uchun alohida $n(n+3)/2$ ta parametr baholanishi kerak, ya’ni jami $l\,n(n+3)/2$ ta parametr; bu kam hajmli tanlanmada amalda bajarib bo‘lmaydigan talabdir.

Ikkinchidan, ko‘p sinfli masalani binar masalalarga keltirish usullari kichik tanlanmada turlicha xulq ko‘rsatadi. «Bittaga qarshi qolganlar» (OvR) sxemasida l ta tasniflagich quriladi, ularning har biri kuchli nomutanosib tanlanmada o‘qitiladi: musbat sinfdagi obyektlar ulushi taxminan $1/l$. «Bittaga qarshi bitta» (OvO) sxemasida $l(l-1)/2$ ta tasniflagich quriladi, ular muvozanatliroq, ammo har biri atigi $m_i + m_j$ ta obyektda o‘qitiladi. Umumlashgan yondashuv - xatolarni tuzatuvchi chiqish kodlari (ECOC) [14, 15]: $M \in \{-1, 0, +1\}^{l\times T}$ kodlash matritsasi tanlanadi, T ta binar tasniflagich f_t o‘qitiladi va qaror

$$A(x) = \underset{k\in Y}{argmin} \sum_{t=1}^{T} L\left(M_{kt}\, f_t(x)\right) \qquad (14)$$

qoidasi bilan qabul qilinadi. OvR va OvO bu sxemaning xususiy hollaridir. R.Rifkin va A.Klautau yaxshi sozlangan regulyarizatsiyalangan binar tasniflagichlar bilan oddiy OvR murakkabroq sxemalardan qolishmasligini ko‘rsatgan [16]; kam hajmli tanlanmada bu xulosa yanada dolzarb, chunki murakkab sxemalarning qo‘shimcha parametrlari ham baholanishi kerak.

Uchinchidan, tibbiy sinflar odatda nomutanosib bo‘ladi: keng tarqalgan tashxislar bilan bir qatorda faqat bir necha obyektdan iborat sinflar uchraydi. Umumiy aniqlik (accuracy) bunday holda kichik sinflardagi xatolarni yashiradi, shuning uchun (15) makro-F1, sinflar bo‘yicha o‘rtacha to‘liqlik yoki ko‘p sinfli AUC (Hand–Till) [21] dan foydalanish kerak. Bu yerda makro-F1 quyidagicha hisoblanadi:

$$F_1^{macro} = \frac{1}{l}\sum_{k=1}^{l} \frac{2\,P_k R_k}{P_k + R_k}, \qquad (15)$$

bunda P_k va R_k - k-sinf bo‘yicha aniqlik (precision) va to‘liqlik (recall).

Kichik sinflar uchun sintetik obyektlar generatsiyasi (SMOTE [22]) kam hajmli tanlanmada ehtiyotkorlik bilan qo‘llanishi lozim: m_k juda kichik bo‘lganda interpolyatsiya mavjud obyektlarni takrorlashdan farq qilmaydi.

To‘rtinchidan, tibbiy sinflar ko‘pincha ierarxik tuzilmaga ega (masalan, XKT-10 tasnifida kasallik guruhi → nozologiya → klinik shakl) va xatoliklar qiymati sinflar orasidagi «masofaga» bog‘liq: bir guruh ichidagi chalkashlik guruhlararo xatodan arzonroq. Bu (4) dagi Λ matritsasining tuzilmasida aks etadi va ierarxiyaning yuqori darajasida qaror qabul qilib, quyi darajada esa rad etish (Δ) imkonini beradi.

Beshinchidan, yondosh kasalliklar mavjudligi (2) dagi kesishmaslik shartini buzadi: bemor bir vaqtda bir nechta sinfga tegishli bo‘lishi mumkin. Bunday holda masala ko‘p belgili tasniflash masalasiga aylanadi va $Y = \{0,1\}^l$ deb olinadi.

Tibbiy ma’lumotlarning tuzilishi: aralash tipli alomatlar muammosi

Tibbiy ma’lumotlar uchun quyidagi alomat tiplari xarakterli: miqdoriy (gemoglobin darajasi, yosh, tana harorati), tartibli (og‘riq intensivligi, kasallik bosqichi, “kam–o‘rta–ko‘p” shkalalari), binar (simptom bor/yo‘q, jins), nominal (qon guruhi, kasallik lokalizatsiyasi). Ushbu tiplar uchun ruxsat etilgan amallar turlicha: nominal alomatlarda faqat tenglikni tekshirish, tartibli alomatlarda - taqqoslash, miqdoriy alomatlarda esa arifmetik amallar ma’noga ega. Bitta formulada ularni farqlamay ishlatish (masalan, nominal kodlarni son sifatida Evklid masofasiga qo‘yish) ma’nosiz natijaga olib keladi.

Aralash tipli alomatlar muammosining uch jihati ajratildi:

1. *Masshtab muammosi.* Miqdoriy alomatlar turli birliklarda o‘lchanadi (masalan, leykotsitlar soni 10^9/l, glyukoza mmol/l), shuning uchun normallashtirilmagan

(cc) BY

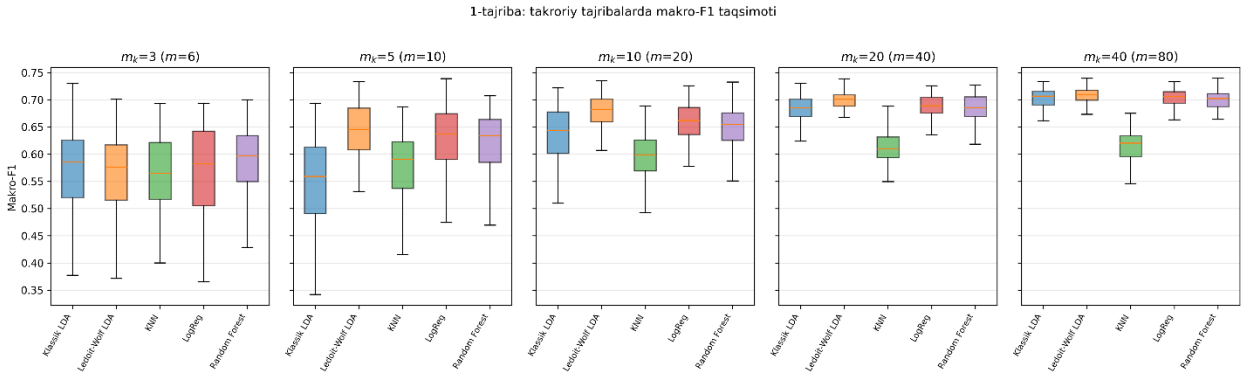
masofada eng katta diapazonli alomat ustunlik qiladi.

2. *Tiplar muammosi.* Binar va nominal alomatlar uchun “farq” tushunchasi miqdoriy alomatlardagidan farq qiladi; nominal alomatlarni “one-hot” kodlash alomatlar sonini oshiradi va (9) shartini kuchaytiradi.
3. *Tushirib qoldirilgan qiymatlar.* Klinik ma’lumotlarda qiymatlarning yo‘qligi ko‘pincha tasodifiy emas (masalan, shifokor tahlilni faqat ma’lum shikoyatda buyuradi). Kichik tanlanmada imputatsiya modelining o‘zi ishonchsiz bo‘lgani uchun (12) dagi kabi yo‘q qiymatlarni o‘tkazib yuboruvchi o‘lchovlar afzalroq [20].

Natijalar

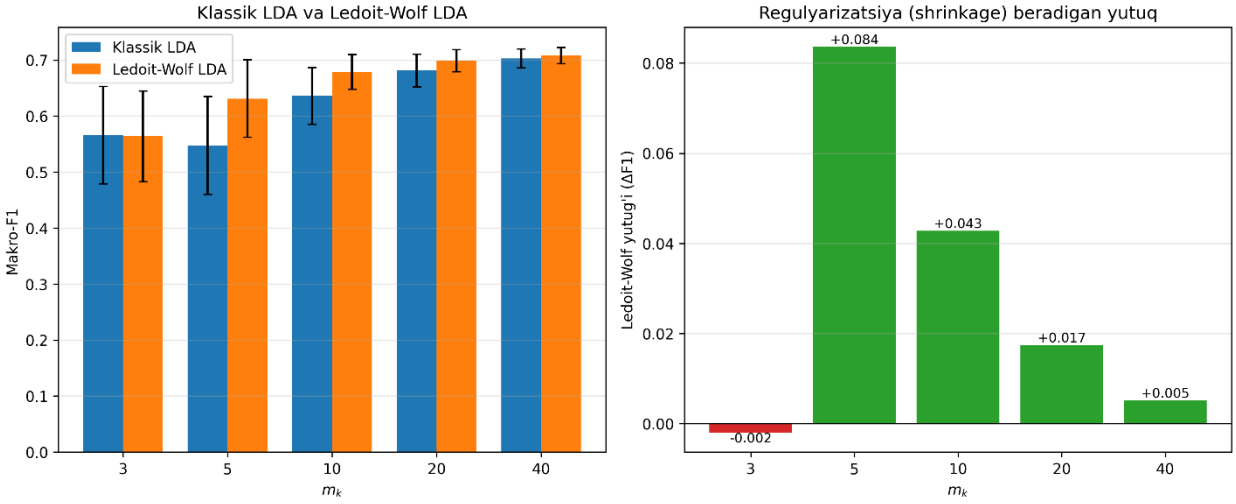
1.1. 1-tajriba. Tanlanma hajmining ta’siri

Bu tajribada ochiq “diabetes” to‘plamidan [23] foydalanildi. Unda 759 ta obyekt, $n = 8$ miqdoriy alomat, $l = 2$ sinf mavjud. Uning tuzilishi kichik klinik tadqiqotlarga xos ma’lumotlarga yaqin. Har bir sinfdan $m_k \in \{3,5,10,20,40\}$ ta obyekt tasodifiy tanlanib o‘qitish tanlanmasi, qolganlari esa nazorat tanlanmasi sifatida olindi va tajriba har bir m_k uchun 100 marta takrorlandi.



1-rasm. Takroriy tajribalarda makro - F1 taqsimoti

Taqqoslangan usullar klassik LDA, Ledoit–Wolf qisqartirishli LDA (11), k eng yaqin qo‘shni (KNN, bu yerda k=1), L_2-regulyarizatsiyalangan logistik regressiya (LogReg) va tasodifiy o‘rmon. Tanlanmadagi alomatlar standartlashtirildi.



2-rasm. Tanlangan usullar yordamida olingan natijalarning taqqoslanishi

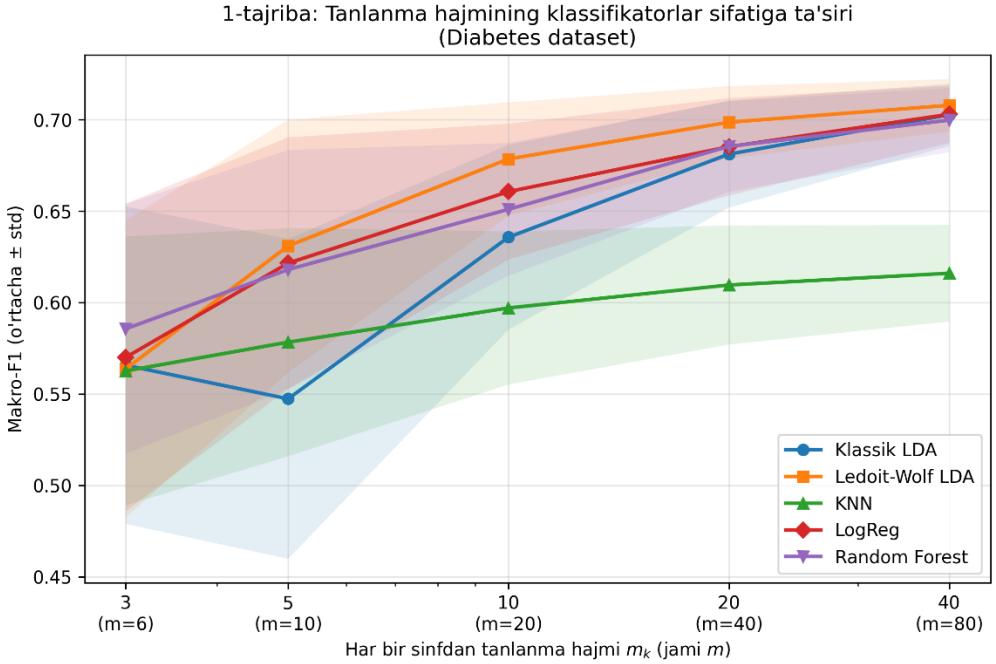
Quyida 2-jadvalda 1-tajriba natijalari keltirilgan.

2-jadval.

“diabetes” tanlanmasida sinfdagi o‘qitish obyektlari soniga bog‘liq ravishda makro-F1 (100 marta takrorlash bo‘yicha o‘rtacha chetlanish)

Tr	m_k	m (jami o‘qitish)	Klassik LDA	LDA (Ledoit–Wolf)	KNN	L_2-logistik regressiya	Tasodifiy o‘rmon (RF)
1	3	6	0,566±0,087	0,564±0,081	0,563±0,073	0,582±0,087	0,591±0,067
2	5	10	0,547±0,087	0,631±0,069	0,571±0,061	0,605±0,070	0,612±0,065
3	10	20	0,636±0,050	0,679±0,031	0,604±0,045	0,669±0,043	0,663±0,037
4	20	40	0,681±0,029	0,699±0,020	0,608±0,034	0,682±0,030	0,681±0,025
5	40	80	0,703±0,017	0,708±0,014	0,618±0,022	0,705±0,017	0,701±0,020

Klassik LDA $m \leq 15$ da eng yomon natija ko‘rsatdi va uning standart chetlanishi boshqa usullarnikidan 2–4 barobar katta bo‘ldi. Qisqartirish (11) qo‘llanganda $m_k = 5$ da makro-F1 0,744 dan 0,934 gacha oshdi. Tanlanma hajmi ortgan sari ($m_k \geq 20$) usullar orasidagi farq standart chetlanish doirasiga tushib qoldi va klassik LDA da eng yaxshi natija olindi. Bu kam hajmli rejimdagi asosiy muammo usulning modeli emas, balki parametrlarni baholash turg’unligi ekanini tasdiqlaydi.

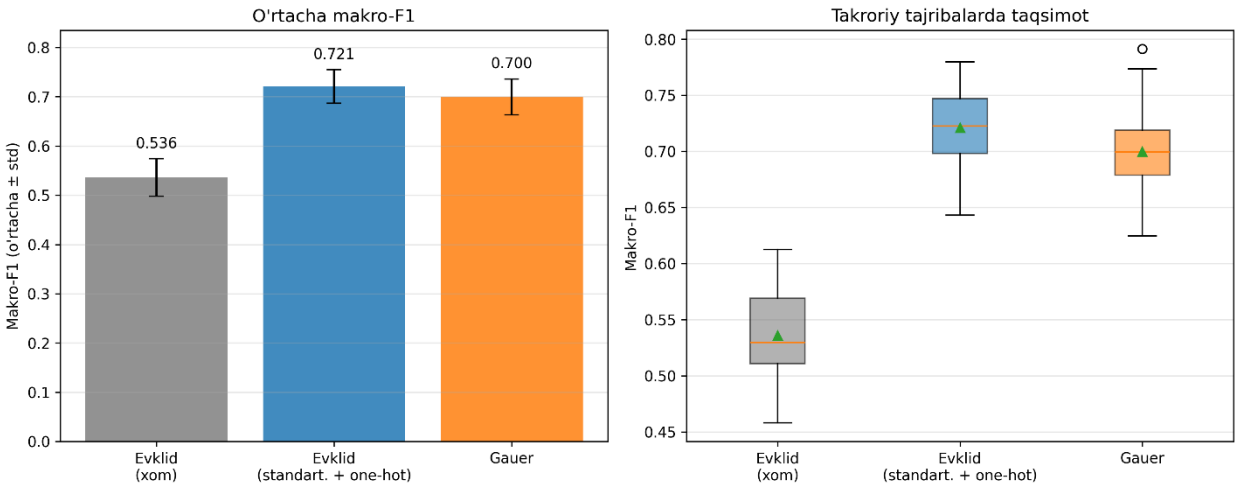


3-rasm. Tanlanma hajmining klassifikatorlar sifatiga ta’siri

1.2. 2-tajriba. Aralash turdagi ma’lumotlar

Bu tajribada to‘rt sinfga ega sintetik tanlanma generatsiya qilindi. Bunda 5 ta turli masshtabli miqdoriy alomat (masshtablar 10^{-1} dan 10^{3} gacha, laborator ko‘rsatkichlarning turli birliklarini imitatsiya qiladi), 3 ta binar, 2 ta besh darajali tartibli va 2 ta uch kategoriyali nominal alomat hosil qilindi. O‘qitish tanlanmasining har bir sinfida 10 tadan obyekt olindi ($m = 40$), tajriba 50 marta takrorlandi. KNN tasniflagichi uchta masofa bilan solishtirildi. Xom ma’lumotlardagi Evklid masofasi standartlashtirish va nominal alomatlarni “one-hot” kodlashdan keyingi Evklid masofasi, Gauer masofasi (12)–(13). Ikkala tajribada ham sifat ko‘rsatkichi sifatida makro-F1 olindi.

(cc) BY



4-rasm. Aralash turdagi ma'lumotlarda KNN uchun turli masofalar (2-tajriba)

Quyida 3-jadvalda 2-tajriba natijalari keltirilgan.

3-jadval.

Aralash toifali sintetik ma'lumotlarda turli masofalar bilan KNN tasniflagichining makro-F1 qiymati ($l = 4$, $m_k = 10$, 50 takrorlash)

T/r	Masofa	Makro_F1_mean	Makro_F1_std
1	Evklid (xom ma'lumotlar)	0.535859928851742	0.038164594003005856
2	Evklid (standartlashtirish + one-hot)	0.7209794580241844	0.033891612255066275
3	Gower	0.6996134733228964	0.036485208821694114

Alomatlar turi va masshtabini hisobga olmaydigan Evklid masofasi tasodifiy taxmin darajasiga yaqin natija berdi. Masofa amalda faqat eng katta masshtabli bitta alomat bilan aniqlandi. Turlarni hisobga oluvchi har ikkala yondashuv sifatni deyarli ikki barobar oshirdi. Bu sintetik tanlanmada standartlashtirish va "one-hot" kodlash Gauer masofasidan bir oz yaxshiroq natija berdi, ammo farq bitta standart chetlanishdan kichik.

4. Muhokama

Ishda olingan natijalar kichik hajmli tibbiy tanlanmalarda tashxis algoritmini tanlashda bir nechta nazariy tamoyilni asoslaydi.

(8) bahosi va 1-tajriba shuni ko'rsatadiki, kichik hajmli tanlanmalarda modelning moslashuvchanligi emas, balki uning parametrlari bahosining turg'unligi hal qiluvchi ahamiyatga ega. Soddalashtirish (11), L_1/L_2 regulyarizatsiya va ansambl usullari bu muammoni turli yo'llar bilan hal qiladi. Bu S. Raudis va A. Jeyn tavsiyalariga [9] hamda klinik bashorat modellari uchun tanlanma hajmini hisoblash bo'yicha zamonaviy tavsiyalarga mos keladi. Shu bilan birga, katta m da klassik LDA ning qisqartirilgan variantdan ustunlik qilishi regulyarizatsiya darajasini tanlanma hajmiga moslab sozlash zarurligini ko'rsatadi.

Sifatni baholash ham kam hajmli tanlanmaga moslashtirilishi kerak. Kichik hajmli tanlanmada test tanlanmasini ajratish va o'qitish uchun berilganlarni yanada kamaytiradi, oddiy kross-validatsiya esa giperparametrlarni tanlash

(cc) BY

bilan birga qo‘llanilganda optimistik siljishga ega bahoni beradi [11]. Shuning uchun ichma-ich (nested) kross-validatsiya yoki ko‘p marta takrorlangan tasodifiy bo‘linishlar qo‘llanishi va natijalar dispersiyasi bilan birga e’lon qilinishi kerak [10]. 2-jadvaldagi standart og’ishlar shuni ko‘rsatadiki, $m_k = 3$ da bitta tasodifiy bo‘linish asosida chiqarilgan xulosa usullarning tartibini butunlay o‘zgartirib yuborishi mumkin.

Kam hajmli ko‘p sinfli masalalarda oddiy dekompozitsiyalar (OvR) va bevosita ko‘p sinfli modellarni (softmax, regulyarizatsiyalangan LDA) tanlash maqul [16] bo’ladi. Juravlyov yondashuvidagi (6) rad etish va (5) qoidadagi λ_0 qiymat klinik amaliyotga bevosita mos keladi. Shuningdek algoritm ishonchli bo’lmagan holatlarni shifokorga natija sifatida qaytaradi. Kichik hajmli tanlanmada ziddiyatli ehtimolliklarning o‘zi ham yomon sozlangab bo‘lishi mumkin, shuning uchun rad etish chegarasini tanlashda ehtiyotkorlik talab etiladi.

2-tajriba shuni ko‘rsatadiki, asosiy yo‘qotish alomatlar turini e’tiborsiz qoldirishdan kelib chiqadi, turlarni hisobga olishning aniq usuli esa ikkinchi darajali ahamiyatga ega. Gauer masofasining afzalligi sifat ko‘rsatkichida emas, balki boshqa alomatlarda namoyon bo‘ladi. U alomatlar sonini oshirmaydi, tushirib qoldirilgan qiymatlar bilan bevosita ishlaydi va har bir alomatning hissasi shifokorga tushunarli bo’ladi. Heterogen masofalarning boshqa variantlari (HEOM, HVDM) ham shu tamoyilga asoslanadi. Daraxtlar ansambllari jadval ko‘rinishidagi ma’lumotlarda yuqori natija ko‘rsatadi [24].

5. Xulosa

Tibbiy tashxis masalalari sun’iy intellekt nuqtai nazaridan beshta mustaqil mezon, ya’ni chiqish tuzilishi, ma’lumot modalligi, bilim manbai, tanlanma hajmi va qaror rejimi bo‘yicha tasniflandi. Bu tasnif aniq klinik masala uchun usullar sinfini tanlashni tizimlashtirish imkoniyatini beradi.

Tashxis masalasining formal qo‘yilishi asimmetrik yo‘qotishlar matritsali bayescha xavf (4)–(5) va rad etish imkoniyatini o‘z ichiga olgan algebraik yondashuv (6) doirasida berilgan. Kam hajmli tanlanma obyektlar, alomatlar va sinflar soni orasidagi munosabatlar (9) yordamida aniqlandi va kovariatsiya matritsasining xosligi sharti $m - l < n$ ekanligi ko‘rsatildi.

Shuningdek ko‘p sinfli tashxis masalalarining beshta o‘ziga xos jihatlari ajratib ko’rsatildi. Sinflarga tegishli bo’lgan ma’lumotlarning kamayishi, dekompozitsiya sxemalarining turlicha jihatlari, sinflar nomutanosibligi, ierarxik tuzilma va yondosh kasalliklar. Ulardan kelib chiqadigan amaliy tavsiyalar asoslandi. Bular makro-metrikalardan foydalanish, sodda dekompozitsiyalarni afzal ko‘rish, rad etish imkoniyatini kiritish hisoblanadi.

Aralash turdagi alomatlar muammosining uch jihati ya’ni masshtab, turlar va tushirib qoldirilgan qiymatlar tahlil qilindi, hamda Gauer o‘xshashlik o‘lchovi (12)–(13) asosida yagona ifodalash taklif etildi.

Adabiyotlar

1. Ledley R. S., Lusted L. B. Reasoning foundations of medical diagnosis // Science. - 1959. - Vol. 130, No. 3366. - P. 9–21.
2. Shortliffe E. H. Computer-Based Medical Consultations: MYCIN. - New York: Elsevier, 1976. - 264 p.
3. Vapnik V. N. Statistical Learning Theory. - New York: Wiley, 1998. - 736 p.
4. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. - 2nd ed. - New York: Springer, 2009. - 745 p.
5. Bishop C. M. Pattern Recognition and Machine Learning. - New York: Springer, 2006. - 738 p.
6. Esteva A., Kuprel B., Novoa R. A. et al. Dermatologist-level classification of skin cancer with deep neural networks // Nature. - 2017. - Vol. 542. - P. 115–118.
7. Rajkomar A., Dean J., Kohane I. Machine learning in medicine // New England Journal of Medicine. - 2019. - Vol. 380, No. 14. - P. 1347–1358.
8. Topol E. J. High-performance medicine: the convergence of human and artificial intelligence // Nature Medicine. - 2019. - Vol. 25. - P. 44–56.
9. Raudys S. J., Jain A. K. Small sample size effects in statistical pattern recognition: recommendations for practitioners // IEEE Transactions on Pattern Analysis and

(cc) BY

Machine Intelligence. - 1991. - Vol. 13, No. 3. - P. 252–264.
10. Vabalas A., Gowen E., Poliakoff E., Casson A. J. Machine learning algorithm validation with a limited sample size // PLoS ONE. - 2019. - Vol. 14, No. 11. - e0224365.
11. Varma S., Simon R. Bias in error estimation when using cross-validation for model selection // BMC Bioinformatics. - 2006. - Vol. 7. - Art. 91.
12. Журавлёв Ю. И. Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. - 1978. - Вып. 33. - С. 5–68.
13. Журавлёв Ю. И., Рязанов В. В., Сенько О. В. «Распознавание». Математические методы. Программная система. Практические применения. - М.: Фазис, 2006. - 176 с.
14. Dietterich T. G., Bakiri G. Solving multiclass learning problems via error-correcting output codes // Journal of Artificial Intelligence Research. - 1995. - Vol. 2. - P. 263–286.
15. Allwein E. L., Schapire R. E., Singer Y. Reducing multiclass to binary: a unifying approach for margin classifiers // Journal of Machine Learning Research. - 2000. - Vol. 1. - P. 113–141.
16. Rifkin R., Klautau A. In defense of one-vs-all classification // Journal of Machine Learning Research. - 2004. - Vol. 5. - P. 101–141.
17. Gower J. C. A general coefficient of similarity and some of its properties // Biometrics. - 1971. - Vol. 27, No. 4. - P. 857–871.
18. Ledoit O., Wolf M. A well-conditioned estimator for large-dimensional covariance matrices // Journal of Multivariate Analysis. - 2004. - Vol. 88, No. 2. - P. 365–411.
19. Tibshirani R. Regression shrinkage and selection via the lasso // Journal of the Royal Statistical Society: Series B. - 1996. - Vol. 58, No. 1. - P. 267–288.
20. Little R. J. A., Rubin D. B. Statistical Analysis with Missing Data. - 3rd ed. - Hoboken: Wiley, 2019. - 462 p.
21. Hand D. J., Till R. J. A simple generalisation of the area under the ROC curve for multiple class classification problems // Machine Learning. - 2001. - Vol. 45, No. 2. - P. 171–186.
22. Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P. SMOTE: synthetic minority over-sampling technique // Journal of Artificial Intelligence Research. - 2002. - Vol. 16. - P. 321–357.
23. Aeberhard S., Coomans D., de Vel O. Comparative analysis of statistical pattern recognition methods in high dimensional settings // Pattern Recognition. - 1994. - Vol. 27, No. 8. - P. 1065–1077.
24. Grinsztajn L., Oyallon E., Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? // Advances in Neural Information Processing Systems 35 (NeurIPS 2022), Datasets and Benchmarks Track. - 2022.

(cc) BY