Кібернетика та комп'ютерні технології

УДК 510.5 DOI:10.34229/2707-451X.23.4.9

В.М. БУДНИК, М.М. БУДНИК

АЛГОРИТМ КЛАСИФІКАЦІЇ ГРУП ПАЦІЄНТІВ НА ОСНОВІ МЕТОДУ ЛІНІЙНОГО ДИСКРИМІНАНТНОГО АНАЛІЗУ

Розроблено алгоритм класифікації декількох груп пацієнтів методом лінійного дискримінантного аналізу (ЛДА), який дозволяє досягти високої точності дискримінації. Його застосовано до аналізу масиву біомедичних даних, отриманих різними діагностичними методами груп здорових осіб та осіб з виявленою хворобою. У результаті застосування цього алгоритму на наборі параметрів отримано дискримінантну функцію та вирішувальне правило, що дозволяє досягти середньої точності 85 %. Крім цього алгоритм застосовано до двох окремих груп параметрів – показників крові та біохімічних аналізів, для яких середня точність дискримінації відповідно досягає 84 % та 90 %.

Ключові слова: *алгоритм, класифікація, аналіз, інформативні параметри, ЛДА.*

Вступ та постановка задачі. Як відомо, основне призначення медицини – це лікування людей. З розвитком медичної галузі [1] пов'язано прискорення виявлення різних травм у організмі, виявлення ознак хвороб, які на ранніх стадіях при застосуванні коректних ліків можна перебороти і навіть повністю вилікувати. Проте з кожним роком у світі завдяки зміні екології та вживанні в їжу генномодифікованих рослин чи тварин відбуваються мутації хвороб, виявлення і лікування яких стає складнішим. У плані лікування тут сприяє розвиток різних галузей наук, у тому числі біології, хімії та різноманітного технологічного обладнання для лабораторій, що допомагає фармацевтам щодня працювати над вдосконаленням формул медичних препаратів. Але в плані виявлення хвороб покращення якості й швидкості діагностики можна досягти за рахунок застосування новіших більш сучасних діагностичних приладів або комплексів.

Розвиток діагностичних приладів відбувається у результаті клопіткої співпраці талановитих науковців, медиків та інженерів. Вони разом відбирають нові діагностичні критерії або комбінують уже відомі та за допомогою статистичних методів перевіряють ефективність їх застосування.

Іноді виникає ситуація, коли медики надають науковцям базу різнопланово досліджених параметрів, і вони мають якимось чином їх поєднати та обробити, щоб з них вибрати параметри, які у подальшому покращать діагностування певної патології або хвороби.

Схожа задача виникла при обробці та класифікації бази даних пацієнтів, наданих ДУ «Інститут педіатрії, акушерства і гінекології ім. О.М. Лук'янової» НАМН України, де потрібно відібрати параметри та отримати вирішувальне правило, яке забезпечувало б високу (>80 %) точність класифікації. Для виконання такої роботи потрібно вирішити наступну задачу: розробити алгоритм підвищення точності дискримінації масиву біомедичних даних, які можуть включати N здорових людей та хворих на різні патології, кожен з яких описується M параметрами. При цьому груп може бути декілька, а кількість пацієнтів та параметрів – досягати сотень.

Застосування відомих алгоритмів класифікації, у тому числі лінійного дискримінантного аналізу (ЛДА) [2], безпосередньо до всієї вибірки даних не дає можливості досягти максимальної точності дискримінації. Тому потрібно розробити спеціальний алгоритм, який буде включати крім етапу класифікації засобами ЛДА, також етап підготовки даних та спеціальний підхід до відбору параметрів.

1. Опис алгоритму

Отже, розроблений алгоритм (рис. 1) включає 5 основних етапів.

1. Попередня обробка – включає 4 підетапи.

1-1. Підготовка масиву даних (необов'язковий).

Підготовка полягає у вибракуванні обстежених, які мали велику кількість пропусків даних та заміни частини пропусків чи анормальних даних (викидів) середніми по вибірках.

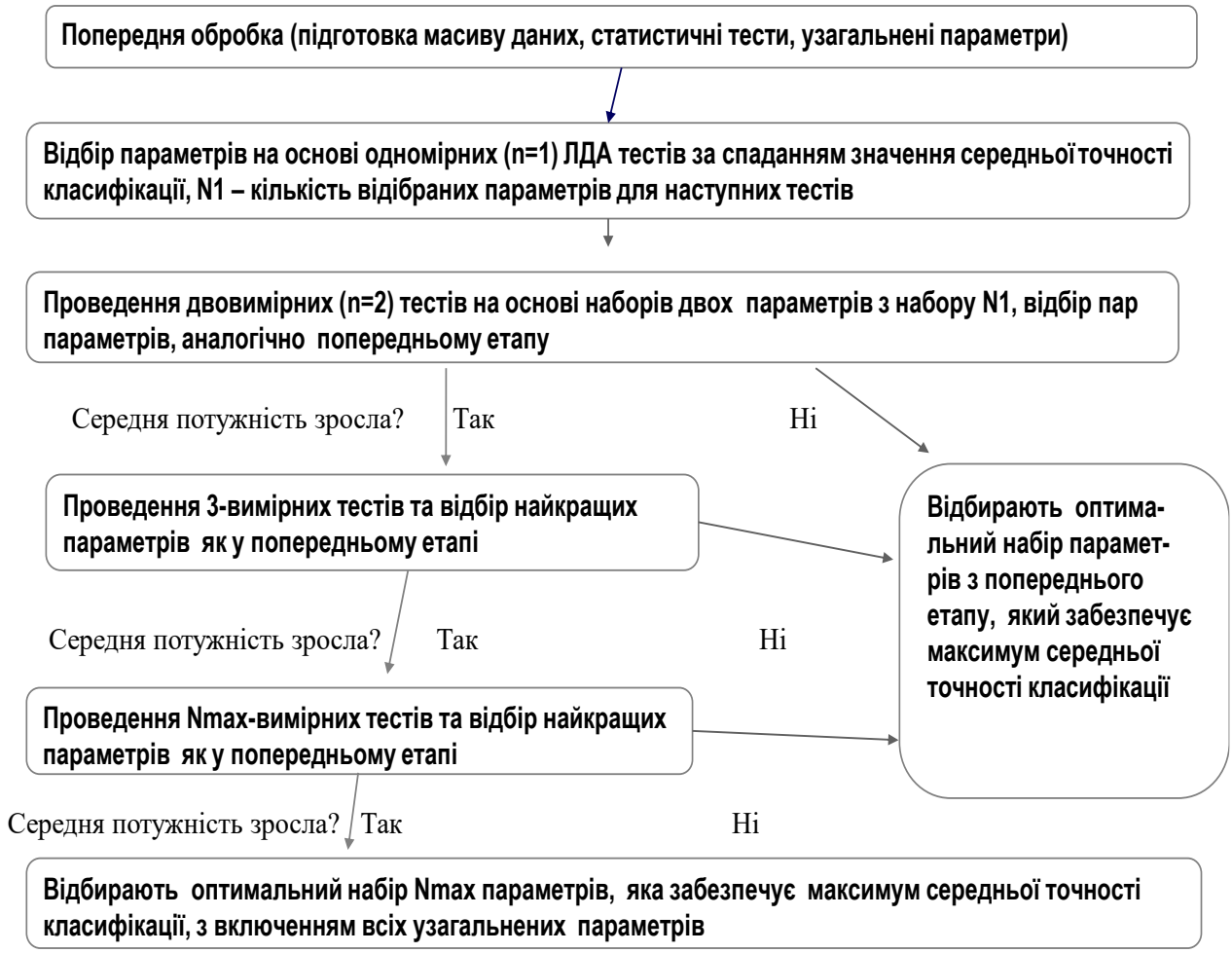

РИС. 1. Блок-схема розробленого алгоритму

1-2. Тести на статистичну відмінність.

Тести на статистичну відмінність включають проведення тестів Стьюдента [3]. Наприклад, для К груп потрібно провести М*К*(К-1)/2 одномірних тестів Стьюдента для кожного з М параметрів між всіма К*(К-1)/2 можливими парами груп. У разі негауссового розподілу вибірок потрібно застосовувати відомі тести на нормальність, наприклад, Колмогорова – Смирнова [4].

1-3. Попередній відбір інформативних параметрів за рівнем значимості тесту. Наприклад, можна застосувати тест Стьюдента та правило, що вибірки статистично відрізняються, якщо $p < 0,05$.

1-4. Пошук узагальнених параметрів (необов'язковий).

Коли аналіз масиву значень параметрів виявив, що деякий набір параметрів споріднений (близький), цей набір доцільно замінити (згорнути) на один узагальнений параметр. Наприклад, у набір можуть входити параметри, які описують результати певного діагностичного методу чи тесту. Також такі узагальнені параметри можуть бути обчислені для певної групи обстежених чи навіть частини групи (підгрупи).

Пошук узагальнених параметрів проводиться аналогічно за етапами 1-1, 1-2, 1-3, 2, 3, 4, 5.

2. Відбір параметрів на основі одномірних ЛДА тестів за зростанням значення середньої точності класифікації.

Проводиться для параметрів, які залишилися, тобто не ввійшли в узагальнені параметри, якщо такі було знайдено на підетапі 1-4.

3. Проведення двовимірних тестів та відбір найкращих параметрів аналогічно попереднього пункту.

4. Проведення багатовимірних тестів з більшою кількістю параметрів та їх відбір, поки потужність буде зростати.

5. Відбір оптимального набору параметрів та обчислення дискримінантної функції, коли середня потужність досягне максимуму.

Проводиться для всіх параметрів, тобто відібраних на основі етапу 4 та усіх узагальнених параметрів (якщо такі є), знайдених на підетапі 1-4.

2. Застосування алгоритму

Даний алгоритм був успішно застосований для вирішення двох задач: для виявлення хвороб сполучної тканини (поділ на 2 групи) та диференціальної діагностики ішемічної хвороби серця (ІХС) та міокардиту (поділ на 4 групи) [5]. В результаті досягнуто збільшення середньої точності класифікації (потужності дискримінації) діагностування хвороб у обох випадках.

Далі розглянемо детальніше застосування алгоритму на основі масиву даних, наданих ДУ «Інститут педіатрії, акушерства і гінекології ім. О.М. Лук'янової» НАМН України, яка містила результати обстежень 501 дитини, обстежених різними діагностичними методами. Завдання полягало у відборі кількісних показників [6], які мають високу діагностичну точність та отримати вирішувальні правила для виявлення дітей з хворобами сполучної тканини [7, 8].

2.1. Відомості про базу даних

Обстежено 295 здорових і 206 хворих дітей. Усім дітям проводилося комплексне лабораторно-інструментальне обстеження, поглиблене біохімічне та імунологічне дослідження з вивченням активності та інтенсивності фагоцитозу, показників НСТ-тесту (спонтанної індукованої активності та функціонального резерву нейтрофільних гранулоцитів), наявності ауто антитіл (антинуклеарних, антифосфоліпідних, анти-ДНК-антитіл) за уніфікованими методиками, що загалом склали 240 параметрів.

2.2. Попередня обробка даних

Перший етап обробки полягає у підготовці масиву даних, що включала відкидання обстежених, які мали велику кількість пропусків даних, заміну середнім арифметичним значенням параметра

пропусків даних та проведення тесту Стьюдента на статистичну відмінність груп. Різницю між показниками вважали достовірною при $p < 0,05$. В результаті було відібрано 24 параметри: g (вміст гамаглобулінів у сироватці крові), CRB (вміст С-реактивний білок у сироватці крові), Limfots (вміст лімфоцитів у крові), MON (абсолютна кількість моноцитів у сироватці крові), TC56 % (відносний вміст цитолітичних Т-клітин (CD56+)), T 25 % (відносний вміст регуляторних (CD25+) Т-лімфоцитів), TR 127 % (відносний вміст активованих (CD 127 +) Т-лімфоцитів), B 5 % (відносний вміст В-лімфоцитів крові), NK 8 % (відносний вміст природних кілерів), TSIKkonts (вміст циркулюючих імунних комплексів у сироватці крові), lgA (вміст імуноглобулінів А у крові), lgM (вміст імуноглобулінів М у крові), lgG (вміст імуноглобулінів G у крові), ANA (титри антинуклеарних антитіл), aDNKnat (титри антитіл до нативної ДНК), aDNKdenat (титри антитіл до денатурованої ДНК), AFLAT aCl (титри антикардоліпінових антитіл), aPs (титри антифосфатидил серинових антитіл), aRe (титри антифосфатидилетаноламінових антитіл), aTSTSP (титри антитіл до циклічного цитрулінованого пептиду), 356 nm, 370 nm, 430 nm та 530 nm (вміст продуктів перекисної модифікації білків, зареєстрованих при даній довжині хвилі в нм).

2.3. Пошук узагальнених параметрів

Наступним етап полягає у виявленні наборів параметрів, які вимірювались блоками для певної підгрупи осіб. В кожній підгрупі пропуски даних замінюємо середнім арифметичним значенням цього параметра. Після цього методом ЛДА генеруємо узагальнені параметри за кожною з підгруп і далі обчислюємо загальну дискримінантну функцію.

Лінійний дискримінантний аналіз (ЛДА) [2] – це один із методів багатовимірного статистичного аналізу. Його суть полягає у тому, щоб на основі значень параметрів об'єкта класифікувати його, тобто віднести до одного з декількох класів деяким оптимальним способом. Критерій оптимальності – це мінімум ймовірності хибної класифікації [7]. ЛДА широко застосовують у медицині [8, 9]. У роботі застосовано пакет STATISTICA ver.13, який забезпечує обчислення дискримінантної функції та потужності дискримінації. При аналізі були вибрані наступні опції:

- Enter independents together (одночасне врахування всіх незалежних змінних), при цьому при обчисленні дискримінантної функції одночасно задіяні всі параметри;
- Unstandardized Function Coefficients (нестандартизовані коефіцієнти функції).

Отже, серед даних після попередньої обробки виявлено три набори параметрів (титри антитіл крові – ANTI, біохімічні показники – Nmsum, вміст продуктів перекисної модифікації білків – TCsum), для кожного з яких було обраховано узагальнений параметр як значення дискримінантної функції, обчисленої для даного набору. та заміни частини пропусків середніми по вибірках.

В результаті отримано такі набори параметрів та групи обстежених:

- ANTI – 7 параметрів (ANA, aDNKnat, aDNKdenat, AFLAT aCl, aPs, aRe, aTSTSP), 65 зд./67 хв;
- NMsum – 4 параметри (356 nm; 370 nm, 430 nm, 530 nm), 146 зд./146 хв;
- TCsum – 5 параметрів (MON; TC56 %, T25 %, TR127 %, NK 8%), 53 зд./58 хв.

Узагальнені параметри мають наступний вигляд:

$$\begin{aligned} \text{ANTI} = {} & -1,016 + 0,497*\text{ANA} + 0,190*\text{aDNKdenat} + 0,082*\text{aTSTSP} - \\ & - 0,034*\text{aPs} + 0,022*\text{aDNKnat} + 0,014*\text{aRe} + 0,009*\text{AFLAT aCl}, \end{aligned} \quad (1)$$

$$\text{NMsum} = -2,304 + 0,731*370\text{nm} + 0,503*530\text{nm} - 0,441*430\text{nm} + 0,160*356\text{nm}, \quad (2)$$

$$\begin{aligned} \text{TCsum} = {} & -3,267 + 0,112*\text{T25 \%} - 0,055*\text{TC56 \%} - 0,050*\text{TR127 \%} - \\ & - 0,005*\text{NK8 \%} + 0,001*\text{MON}. \end{aligned} \quad (3)$$

Ймовірність правильної класифікації здорових, хворих та точність дискримінації за цими узагальненими параметрами показана у таблиці.

ТАБЛИЦЯ. Точність дискримінації по узагальнених параметрах

№	Узагальнений параметр	Точність дискримінації, %		
		Здорові	Хворі	Середня
1	ANTI	93,8	57,6	75,6
2	NMsum	72,6	61,6	67,1
3	TCsum	62,3	74,1	68,5
	Середнє	76,2	64,4	70,4

Отже, як було описано вище з 24 параметрів було відібрано 16, з яких отримано 3 узагальнених показники: ANTI, NMsum, TCsum, які включають відповідно 7, 4 та 5 параметрів. Проте з таблиці видно, що по узагальненим параметрам середня точність рівна 70 % (здорові 76 %, хворі 64 %), що недостатньо для надійної діагностики вказаних хвороб.

3. Обчислення дискримінантних функцій, які забезпечують максимум точності

3.1. Класифікація на основі загального набору параметрів

На основі розробленої методики методом ЛДА було проаналізовано 24 – (7 + 4 + 5) = 8 параметрів, що залишились, та оцінена потужність дискримінації на основі кожного з них. В результаті кожна особа описувалась 11 параметрами: g, CRB, Limfots, B5 %, TCsum (3), TSIKkonts, lgA, lgM, lgG, ANTI (1), NMsum (2). Після вибракування частини осіб та заміни пропусків даних середніми по вибірках отримано групи: 75 здорових і 84 хворих. Дискримінантна функція описується наступною формулою:

$$\begin{aligned} D = {} & -2{,}539 + 3{,}305*\text{TSIKkonts} + 0{,}628*\text{CRB} - 0{,}304*\text{TCsum} + \\ & + 0{,}171*\text{g} - 0{,}091*\text{B5 \%} - 0{,}065*\text{lgG} + 0{,}046*\text{ANTI} + \\ & + 0{,}041*\text{NMsum} - 0{,}040*\text{Limfots} + 0{,}025*\text{lgA} - 0{,}012*\text{lgM}. \end{aligned} \quad (4)$$

Вирішувальне правило для віднесення особи до певної групи має наступний вигляд:

$$\text{Якщо } D > 0 \text{ – хворий, якщо } D < 0 \text{ – здоровий.} \quad (5)$$

При цьому, ймовірність правильної класифікації (потужність дискримінації) здорових становить 86,7 %, хворих – 83,3 % (середня рівна 85 %), що достатньо для надійної діагностики хвороб сполучної тканини у дітей.

3.2. Класифікація на основі параметрів аналізу крові

Початкові групи: 0 – здорові (295 осіб) і 1 – хворі (206 осіб). Кожна група описується 51 параметром. Після вибракування пропусків та заміни частини пропусків середніми по вибірках отримано групи: 33 здорових і 48 хворих, що описуються 30-ма параметрами. В результаті аналізу відібрано 16 інформативних параметрів. Дискримінантна функція описується формулою:

$$\begin{aligned} D = {} & 5{,}712 + 4{,}491*\text{TSIKkonts} - 0{,}375*\text{Monots} + 0{,}310*\text{IRI} + 0{,}249*\text{ANA} - \\ & - 0{,}172*\text{lgM} + 0{,}122*\text{lgG} + 0{,}095*\text{Leykots} - 0{,}092*\text{Segment} + \\ & + 0{,}087*\text{var8} + 0{,}062*\text{Eritr} - 0{,}051*\text{var4} + 0{,}038*\text{lgA} + 0{,}030*\text{NK \%} - \\ & - 0{,}019*\text{TA \%} - 0{,}012*\text{Hb} - 0{,}011*\text{Limfots}, \end{aligned} \quad (6)$$

де TSIKkonts – вміст циркулюючих імунних комплексів; Monots – абсолютний моноцитоз; IRI – значення імунорегуляторного індексу; ANA – титри антинуклеарних антитіл; lgM – вміст імуноглобуліну M; lgG – вміст імуноглобуліну G; Leykots – показник лейкоцитозу (в мкл); Segment – відносний вміст сегментоядрових лімфоцитів; var8 – відносний вміст цитотоксичних Т-лімфоцитів; Eritr – кількість еритроцитів в мкл крові; var4 – показник аналізу крові; lgA – вміст імуноглобуліну A;

NK % – відносний вміст природних кліерів; TA % – відносний вміст активованих Т-лімфоцитів; Hb – вміст гемоглобуліну в сироватці крові; Limfots – відносний вміст лімфоцитів у сироватці крові.

Аналогічно застосовуємо вирішувальне правило (5) та отримуємо потужність дискримінації здорових – 78,8 %, хворих – 89,6 %, а середня дорівнює 84,2 %.

3.3. Класифікація на основі біохімічних параметрів

Початкові групи: 0 – здорові (291 осіб) і 1 – хворі (206 осіб). Кожна група описувалась 21 параметром. Після вибракування пропусків вимірювань та заміни частини пропусків середніми по вибірках отримано групи: 44 здорових і 45 хворих, що описуються 14 параметрами. В результаті аналізу відібрано 11 інформативних параметрів. Дискримінантна функція має вигляд:

$$\begin{aligned} D = {} & -10{,}759 - 1{,}884*530\text{nm} + 0{,}424*430\text{nm} - 0{,}384*370\text{nm} - 0{,}314*356\text{nm} + \\ & + 0{,}262*g + 0{,}227*a2 + 0{,}209*b + 0{,}061*\text{Albuminy} - 0{,}056*\text{Timol} + \\ & + 0{,}049*a1 + 0{,}026*\text{CRB}, \end{aligned} \tag{7}$$

де 530 nm – вміст продуктів перекисної модифікації білків (ПМБ) при довжині хвилі 530 nm; 430 nm – вміст продуктів ПМБ при довжині хвилі 430 nm; 370 nm – вміст продуктів ПМБ при довжині хвилі – 370 nm; 356 nm – вміст продуктів ПМБ при довжині хвилі 356 nm; g – вміст γ-глобулінів; a2 – вміст α2-глобулінів; b – вміст β-глобулінів; Albuminy – вміст альбумінів; Timol – значення показнику тимолової проби; a1 – вміст α1-глобулінів; CRB – вміст С-реактивного білка (СРБ).

Аналогічно застосовуємо вирішувальне правило (5) та отримуємо точність дискримінації здорових – 86,6 %, хворих – 93,3 %, а середня дорівнює 90 %.

Висновки

У роботі запропоновано алгоритм класифікації даних на групи, описано особливості попередньої обробки масиву даних, відбору параметрів та їх аналізу. Крім цього показано результати застосування цього алгоритму до масиву даних пацієнтів для задачі виявлення дітей з хворобами сполучної тканини. При цьому проведено аналіз даних 501 обстежених (295 здорових, 206 хворих), кожен з яких описувався 240 параметрами. В результаті обробки з 240 параметрів за t-тестом Стьюдента було відібрано 24. З 16 параметрів методом ЛДА отримано 3 узагальнених показники ANTI, NMsum, TCsum (1–3), які включають відповідно 7, 4 та 5 параметрів. На основі загального набору 8 параметрів та 3 узагальнених отримана кінцева дискримінантна функція, яка забезпечила середню потужність дискримінації 85 %, на основі параметрів аналізу крові – 85,2 %, а класифікація на основі біохімічних параметрів – 91 %, що достатньо для надійної діагностики хвороб сполучної тканини у дітей [10].

Подяки. Автори приносять подяку Національній академії наук України за підтримку в рамках виконання тем відділу 220 ВФ.220.29 «Розробити методи і алгоритми обробки сигналів різної природи для побудови діагностично-інформаційних систем» (2019–2023) та ВП 220.28 «Розроблення нових алгоритмів і програмно-інструментальних засобів для дослідження слабких магнітних сигналів в біології» (2018–2023), (керівник обох М.А. Прімін). Також автори висловлюють подяку д.м.н. О.А. Ошлянській з ДУ «Інститут педіатрії, акушерства і гінекології ім. О.М. Лук'янової» НАМН України за наданий масив даних та постановку задачі.

Список літератури

1. Медицина та охорона здоров'я новітньої доби на теренах України. https://vue.gov.ua/Медицина_та_охорона_здоров'я_новітньої_доби_на_теренах_України (звернення: 04.09.2023)
2. Лінійний дискримінантний аналіз. https://uk.wikipedia.org/wiki/Лінійний_розділювальний_аналіз (звернення: 04.09.2023)

3. Ромакін В.В. Комп'ютерний аналіз даних. Навчальний посібник. Миколаїв: Вид-во МДГУ ім. Петра Могили, 2006. 144 с. https://lib.chmnu.edu.ua/index.php?m=2&b=210
4. Критерій узгодженості Колмогорова – Смирнова. https://uk.wikipedia.org/wiki/Критерій_узгодженості_Колмогорова (звернення: 04.09.2023)
5. Будник В.М., Ошлянська О.А., Будник М.М. Виявлення інформативних показників для діагностики захворювань сполучної тканини у дітей методом ЛДА. Тези доповідей III міжнародної науково-практичної конференції «Інтелектуальні системи в промисловості і освіті» (ІСПО-2011). Суми, 2–4.11.2011. Суми: Видавництво СумДУ. 2011. С. 19–23. https://essuir.sumdu.edu.ua/bitstream-download/123456789/24955/1/Budnik.pdf
6. Сосницька Т.В., Будник В.М., Стаднюк Л.А., Моїсеєнко Є.В., Сосницький В.М., Єгорова Л.В., Лілякевич А.А. Аналіз відтворюваності магнітокардіографічних параметрів. *Український кардіологічний журнал*. 2012. № 2. С. 81–83. https://drive.google.com/file/d/1S5epkuP0TH4p6TwLj0qfZHaOc3O04CAo/view?usp=sharing
7. Ошлянська О.А., Омельченко В.П., Чернишов В.П., Галазюк Л.В. Роль неспецифічної клітинної імунної відповіді у формуванні аутоімунітету. *Перинатологія і педіатрія*. 2008. № 3. С. 83–85.
8. Ошлянська О.А. Маркери деструкції сполучної тканини при вроджених та набутих її патологіях у дітей. *Перинатологія і педіатрія*. 2009. № 4. С. 57–61.
9. Чернишева Д.С., Будник М.М. Застосування дискримінантного аналізу до обробки МКГ інформації. *Комп'ютерні засоби, системи та мережі.* 2004. № 3. С. 57–64. http://dspace.nbuv.gov.ua/handle/123456789/6406
10. Будник В.М. Дослідження і сертифікація біомедичних інформаційно-вимірювальних систем: дисертація канд. тех. наук: 05.13.06. Київ, 2021. 270 с. http://irbis-nbuv.gov.ua/aref/0421U104068

Одержано 20.09.2023

Будник Віталій Миколайович,
кандидат технічних наук, науковий співробітник
Інституту кібернетики імені В.М. Глушкова НАН України, Київ,
https://orcid.org/0000-0001-6296-4065
vitaliy.budnyk@gmail.com

Будник Микола Миколайович,
доктор технічних наук, головний науковий співробітник
Інституту кібернетики імені В.М. Глушкова НАН України, Київ,
Київський національний університет імені Тараса Шевченка,
Сумський державний університет.
http://orcid.org/0000-0002-4020-0213
budnykmykola@gmail.com

УДК 510.5

В.М. Будник [1 *], М.М. Будник [1, 2, 3]

Алгоритм класифікації груп пацієнтів на основі методу лінійного дискримінантного аналізу

[1] *Інститут кібернетики імені В.М. Глушкова НАН України, Київ*
[2] *Київський національний університет імені Тараса Шевченка*
[3] *Сумський державний університет*
[*] *Листування: vitaliy.budnyk@gmail.com*

Вступ. Стаття присвячена розробленню алгоритму класифікації декількох груп пацієнтів з використанням методу ЛДА. Алгоритм дозволяє досягти високої потужності дискримінації, працювати з різнотипними масивами параметрів та розділяти більш ніж дві групи пацієнтів. Його застосовано при проведенні аналізу біомедичних даних, а саме до масиву показників різноманітної природи, які описують здорових осіб та осіб з виявленою хворобою.

Мета роботи. Полягає у розробці алгоритму підвищення потужності дискримінації масиву медичних даних, які включають сотні здорових людей та хворих на хвороби сполучної тканини, а кожна людина описується сотнями параметрів.

Результати. Авторами запропоновано алгоритм підвищення точності класифікації на основі відбору параметрів за допомогою послідовних багатовимірних ЛДА тестів. Наведено приклад застосування алгоритму до задачі аналізу досить великого масиву даних (501 особа: 295 здорових, 206 хворих, 240 параметрів) та виявлення інформативних параметрів для діагностики дітей з хворобами сполучної тканини. У результаті застосування цього алгоритму отримано дискримінантну функцію та вирішувальне правило, яке дозволяє досягти середньої точності дискримінації по всьому набору параметрів у 85 %. Крім цього алгоритм застосовано до двох окремих груп параметрів – показників крові та біохімічного аналізу, при цьому середня точність дискримінації відповідно досягає 84 % та 90 %.

Висновки. Розроблено алгоритм класифікації груп пацієнтів з застосуванням ЛДА, який дозволяє досягти високої точності дискримінації, наведено його структуру та основні етапи. Результати його застосування до вирішення реальної задачі класифікації медичних даних показують його здатність підвищити точність класифікації.

Ключові слова: алгоритм, класифікація, аналіз, інформативні параметри, ЛДА.

MSC 68M15, 68N30, 68U35

Vitalii Budnyk [1 *], Mykola Budnyk [1, 2, 3]

Algorithm for Classification of Patient Groups Based on the LDA Method

[1] *V.M. Glushkov Institute of Cybernetics of the NAS of Ukraine, Kyiv*
[2] *Taras Shevchenko National University of Kyiv*
[3] *Sumy State University*
[*] *Correspondence: vitaliy.budnyk@gmail.com*

Introduction. The article is devoted to the development of the algorithm of classification of several groups of patients using the LDA method at certain stages of its work. The algorithm allows achieve high discrimination value, to work with a variety of parameters and to separate more than two groups of patients. It is used for the analysis of biomedical data, namely for the array of indicators of various nature described healthy persons and persons with the detected disease.

The purpose of paper is to develop an algorithm to increase the power of discrimination of an array of medical data that may include hundreds of healthy people and patients with connective tissue diseases, and each person is described by hundreds of parameters.

Results. The authors proposed an algorithm for increasing classification accuracy based on parameter selection using sequential multivariate LDA tests. An example of the application of the algorithm to the task of analyzing a fairly large array of data (501 individuals: 295 healthy, 206 patients, 240 parameters) and identifying informative parameters for diagnosing children with connective tissue diseases is given. As a result of the application of this algorithm, a discriminant function and a decision rule were obtained, which allows achieve an average accuracy of discrimination over the entire set of parameters of 85 %. In addition, the algorithm is applied for two separate groups of parameters - blood indicators and biochemical analysis, while the average accuracy of discrimination reaches 84 % and 90 %, respectively.

Conclusions. The algorithm of classification groups of patients with the use of LDA has been developed, which allows achieve high accuracy of discrimination. The results of its application in solving the real data set are given. The results of its application to solving the task of classification of medical data show its ability to improve classification accuracy.

Keywords: algorithm, classification, analysis, informative parameters, LDA.