

М. М. Савула, В. К. Паучок, Н. С. Кравченко

ЗАСТОСУВАННЯ МОДИФІКОВАНОГО КЛАСТЕРНОГО АНАЛІЗУ ДАНИХ СТАНДАРТНОГО МЕДИЧНОГО ОБСТЕЖЕННЯ ДЛЯ ДИФЕРЕНЦІАЛЬНОЇ ДІАГНОСТИКИ ДЕЯКИХ ЗАХВОРЮВАНЬ ЛЕГЕНЬ

Тернопільський державний медичний університет ім. І. Горбачевського
Тернопільський національний економічний університет

Встановити етіологію захворювання легень за даними первинного медичного обстеження нелегко. Це зумовлено схожістю клінічних симптомів при різних хворобах, близькістю більшості результатів стандартних лабораторних і рентгенологічних досліджень (за їхніми якісними та кількісними характеристиками), їх непатогномонічністю. По суті лише характерні цитологічні або гістологічні зміни (при раку і саркоїдозі), або бактеріологічне підтвердження (при туберкульозі) дають змогу встановити діагноз. При неспецифічних пневмоніях таким підтвердженням часто є лише результат тест-терапії. Для удосконалення діагностики пропонуються диференціально-діагностичні таблиці, алгоритми, бальна оцінка симптомів та інші [4].

Метою дослідження є вивчення придатності для діагностичного процесу математичного кластерного аналізу багатовимірних даних. Такий метод в окремих випадках використовувався при розв'язанні діагностичних проблем в медицині [1, 3].

Матеріали і методи дослідження

Використано дані обстеження 141 хворого. Рак легень встановлено у 35 (24,82 %) хворих, туберкульоз легень — у 29 (20,56%), саркоїдоз — у 49 (29,8 %); затяжні пневмонії — у 35 (24,82 %). Виходили з припущення, що набір якісних і кількісних характеристик ознак, що стосуються статі, віку хворих, функції зовнішнього дихання (ФЗД), результатів простих лабораторних і загальнодоступних рентгенологічних досліджень (крім даних традиційних обстежень у масив даних включені результати визначення мінеральної щільності кісткової тканини (МШКТ) поперекового відділу хребта, які виконували всім пацієнтам в рамках планового наукового дослідження), вказує на точку в багатовимірному просторі значень цих величин.

Сподівалися, що для пацієнтів з різними захворюваннями вони будуть згруповані в дещо різних областях такого багатомірного простору, тобто окремим недугам відповідатимуть різні підобласті — кластери.

Для розподілу масиву даних на кластери застосовано вдосконалену процедуру ідентифікації ступеню належності точки до розмитої (нечіткої) множини [2] методом «звисяючого грона», вперше запропонованого Д. Бездеком у 1981 році [5]. За допомогою цього методу не вдалося чітко розділити наявні дані на групи для окремих захворювань, тому в нього внесено вдосконалення, описане нижче.

Результати дослідження та їх обговорення

Щоб вияснити придатність кластерного аналізу для вирішення поставленої мети виконано описаний нижче обчислювальний експеримент.

Для розгляду взято показників обстеження хворих. Перелік цих показників подано в табл. 1. Їх значення відомі для $m = 141$ пацієнтів, з різними захворюваннями: рак (позначимо його індексом r), саркоїдоз (s), туберкульоз (t), затяжна пневмонія (z). Множину всіх груп захворювання позначимо символом $J = \{r, s, t, z\}$.

Позначимо результати обстеження одного хворого

$$x_{ik}^j, i = \overline{1, n}; k = \overline{1, m}; j \in J, \quad (1)$$

x_{ik}^j — значення i -того показника для k -того хворого на j -ту хворобу, i — номер показника; k — номер пацієнта; j — індекс захворювання, m — кількість пацієнтів; n — кількість показників; J — множина індексів захворювань.

Таблиця 1

Перелік використаних показників

№	Назва показника	№	Назва показника
1	Вік, роки	Рентгенологічні зміни	
2	Стать, жіноча (0), чоловіча (1)	15	В одній легені (0), в двох (1)
3	Маса тіла, кг.	16	Вогнища, нема (0), є (1)
4	Початок поступовий (0), гострий (1)	17	Інфільтрати, нема (0), є (1)
5	Тривалість хвороби до звернення, міс	18	Деструкції, нема (0), є (1)
6	Куріння, ні (0), так (1)	19	Розмір деструкції, см
7	Супровідні хвороби, нема (0), є (1)	20	Збільш. лімфовузлів, нема (0), є (1)
8	Артралгії, нема (0), є (1)	Показники ФЗД	
Клінічний аналіз крові		21	ЖЕЛ, %
9	ШОЕ, мм/год	22	ОФВ ₁ , %
10	Лейкоцити, 10^9 / л	23	МВЛ, %
11	Лімфоцити, %	МШКТ	
12	Паличкоядерні гранулоцити, %	24	L1-2, %
13	МБТ в мазку, «-» (0), «+» (1)	25	L4, %
14	Проба Манту, мм		

Кількість хворих на окремі недуги позначимо m_r (рак легень), m_t (туберкульоз легень), m_s (саркоїдоз), m_z (затяжні пневмонії). Одна величина $X_k^j = \{x_{ik}^j, i = 1, n\}$ відображає точку у n -вимірному просторі ($k = 1, m; j = r, s, t, z$). Необхідно розділити об'єднання множин

$$X = \{X_k^j; k = \overline{1, m}; j \in J\} \quad (2)$$

на групи, які відповідають окремими захворюванням.

Для цього здійснене нелінійне перетворення координат (1) так, щоб на їх основі отримати дійсні числа

$$y_{ik}^j, i = \overline{1, n}; k = \overline{1, m}; j \in J \quad (3)$$

у співмірному масштабі, які відповідають такій умові: область значень величини (2) має бути тим більше розширена, чим важливіше значення вона має з погляду діагностування. Дотримання такої умови дає підстави застосувати до даних (3) методи кластерного розділення, які ґрунтуються на лінійній метриці. Нижче описано метод такого перетворення.

Розглянемо спочатку показники з цілими (логічними) значеннями. Показники з індексами $i \in I_L$, де $I_L = \{2, 4, 6, \dots, 8, 13, 15, \dots, 18, 19\}$ приймають значення 0 або 1. Вони відображають наявність ознаки хвороби. В межах j -тої нозологічної групи ймовірність цієї ознаки дорівнює її математичному сподіванню:

$$p_i^j = \sum_{k=1}^{m_j} x_{ik}^j / m_j; i \in I_L; j \in J \quad (4)$$

Величина

$$y_{ik}^j = x_{ik}^j p_i^j; k = \overline{1, m}; i \in I_L; j \in J \quad (5)$$

належить відріzk i відображає положення k -тої точки на i -тій осі в j -тій групі хворих.

Показники x_{ik}^j з індексами $i \in I_D$, де $I_D = \{1, 3, 9, \dots, 12, 14, 21, \dots, 25\}$ — це результати вимірювань, відображені дійсними числами. Вони належать області значень $\Omega_i^j = [\min(x_{ik}^j), \max(x_{ik}^j)]$. Позначимо їх можливі значення символом $x_i^j, x_i^j \in \Omega_i^j (i \in I_D; j \in J)$. Зауважимо, що в областях Ω_i^j є підобласті, суттєві і несуттєві з погляду діагностування. Величинам x_{ik}^j відповідають закони розподілу $\rho_i^j(x_i^j)$. Їх легко встановити на основі даних (1). За розподілом $\rho_i^j(x_i^j) (i \in I_D; j \in J)$ знаходимо густину ймовірності величини $x_{ik}^j \in \Omega_i^j$:

$$P_i^j(x_i^j) = \rho_i^j(x_i^j) \left[\int_{\min(x_{ik}^j)}^{\max(x_{ik}^j)} \rho_i^j(x_i^j) dx_i^j \right]^{-1}; (x_i^j \in \Omega_i^j). \quad (6)$$

Помноживши

$$\tilde{y}_{ik}^j = x_{ik}^j \left[1 + P_i^j(x_{ik}^j) \right]; k = \overline{1, m}; i \in I_D; j \in J, \quad (7)$$

отримуємо величини $\tilde{y}_{ik}^j$, які збільшені відносно x_{ik}^j на їх ймовірність. Тобто, значення y_{ik}^j , яке належить відріzk $\left[\min_k(\tilde{y}_{ik}^j), \max_k(\tilde{y}_{ik}^j) \right]$, відображає положення k -тої на i -тій осі в j -тій групі хворих так, що підобласті цього відрізка, важливі з погляду діагностування, мають більший розмір, ніж неважливі підобласті. За міру цієї важливості взято густину ймовірності. Нормуючи $\tilde{y}_{ik}^j$ на одиницю

$$y_{ik}^j = \tilde{y}_{ik}^j \left[\max_k \tilde{y}_{ik}^j \right]^{-1}; k = \overline{1, m}; i \in I_D; j \in J, \quad (8)$$

отримуємо величину, яка належить відріzk.

Формули (4, 5) дають значення величини $y_{ik}^j (k = \overline{1, m}; j \in J)$ для показників $i \in I_L$. Формули (6–9) дають значення цих величин для показників $i \in I_D$. За допомогою названої вище програмної функції системи Matlab величини

(3) розділено на окремі області, що відповідає розв'язанню поставленої задачі. Результати цього експерименту показано в табл. 2.

Як видно з таблиці 2 в результаті обчислень досліджувані дані розділено не на 4 групи, як сподівалися по кількості захворювань, а на 3 групи. В 1-у групу віднесено 94.29 % і 71.43 % від загальної кількості хворих відповідно на рак і на пневмонії (а також — 3.45 % від хворих на туберкульоз). В 2-гу групу віднесено 100.00 % від хворих на саркоїдоз і 25.7 % від хворих на рак легень. В 3-тю групу 96.55 % від хворих на туберкульоз і 28.57 % від хворих на пневмонії.

Зауважимо, що 71.43 % від хворих на пневмонії, віднесено в групу, де переважали хворі на рак легень і 28.57 % від них віднесено в групу, де переважали хворі на туберкульоз. Це може бути пов'язано з тим, що нерідко неспецифічне запалення супроводжує і рак, і туберкульоз. Цей обчислювальний експеримент ілюструє успішне застосування розробленого методу для розділення даних попереднього обстеження хворих із легеневиими хворобами.

Таким чином, отримавши відповідь, що хворий, якому необхідно встановити діагноз, віднесений у 1 групу, існує велика ймовірність раку легень і дещо менша — затяжної пневмонії. Тому пацієнтові необхідно терміново провести більш інформативні дослідження (бронхоскопію, цитологічне дослідження, комп'ютерну томографію та ін.), не нехтуючи інтенсивною терапією антибіотиками широкого спектру дії. Саркоїдоз, очевидно, можна виключити, ймовірність туберкульозу невелика, але це не виключає обов'язкового пошуку мікобактерій.

У хворого, якого віднесено до 2-ї групи найбільш ймовірний саркоїдоз, хоча не можна виключати рак легень, тому йому необхідно згадати додаткові обстеження. Туберкульоз і затяжні пневмонії, очевидно, можна виключити.

У пацієнта 3-ї групи найбільш ймовірний туберкульоз, хоча можлива пневмонія. Йому показана інтенсивна терапія антибіо-

Таблиця 2

Результат розділення отриманих даних (1) на кластери

Кількість хворих в 1 групі	59	41.84 % від загальної кількості хворих	
Рак легень	33	55.93 % від хворих в 1 гр.	94.29 % від хворих на рак легень
Туберкульоз	1	1.69 % від хворих в 1 гр.	3.45 % від хворих на туберкульоз
Саркоїдоз	0		0
Пневмонія	25	42.37 % від хворих в 1 гр.	71.43 % від хворих на пневмонії
Кількість хворих в 2 групі	44	31.20 % від загальної кількості хворих	
Рак легень	2	18.7 % від хворих в 2 гр.	25.7 % від хворих на рак легень
Туберкульоз	0		0
Саркоїдоз	42	95.45 % від хворих в 3 гр.	100.00 % від хворих на саркоїдоз
Пневмонія	0		0
Кількість хворих в 3 групі	38	26.95 % від загальної кількості хворих	
Рак легень	0		0
Туберкульоз	28	73.68 % від хворих в 3 гр.	96.55 % від хворих на туберкульоз
Саркоїдоз	0		0
Пневмонія	10	26.32 % від хворих в 3 гр.	28.57 % від хворих на пневмонії

тиками, виключивши препарати з протитуберкульозною активністю і, очевидно, багаторазові дослідження харкотиння на МБТ всіма доступними методами.

Утворення характерних скупчень на площині чисельних значень використаних показників (2 показники на кожній діаграмі) в різних групах хворих показано для прикладу на рис. 1 і 2. Рисунки ілюструють можливості диференціації певних захворювань на підставі кластерного аналізу комплексу простих обстежень, які можна виконати в амбулаторних умовах.

Враховуючи, що по суті немає методу дослідження, який дозволяв би в 100 % випадків встановити точний діагноз (помилка можлива як на рівні виконання, так і на рівні інтерпретації), слід вважати, що запропонований метод може надати вагомий допис лікарю при встановленні попереднього діагнозу.

Запропонований метод дозволяє уникнути суб'єктивізму, за його допомогою можна швидко врахувати велику кількість показників. Вперше для кластерного аналізу даних медичного обстеження з метою диференціальної діагностики застосовано нелінійні перетворення систем координат, що дозволило підвищити інформативність діагностичного методу.

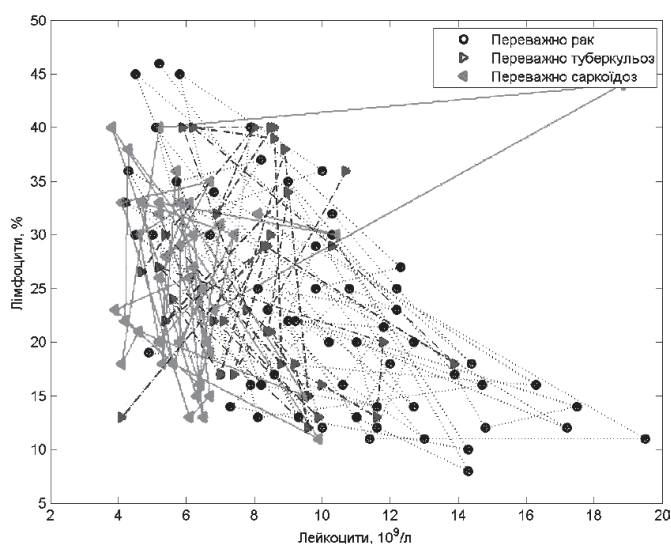


Рис. 1. Діаграма розділення значень показників лімфоцити-лейкоцити.

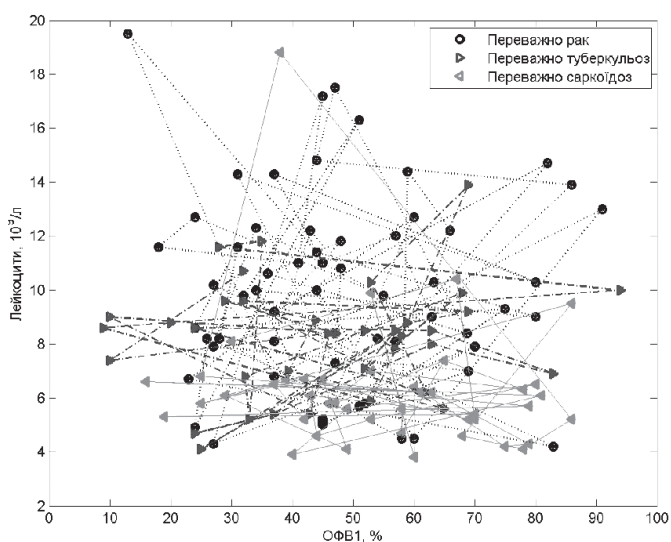


Рис. 2. Діаграма розділення значень показників лейкоцити-ОФВ₁.

Висновки

1. Запропонований метод кластерного розподілу даних первинного медичного обстеження з нелінійним перетворенням їх координат дає якісні результати щодо розрізнення групи хворих на рак, саркоїдоз і туберкульоз. Для такого ж розрізнення затяжної пневмонії взятих даних недостатньо.

2. Метод може слугувати для додаткової оцінки результатів первинного медичного обстеження хворих з легеневиими недугами при встановленні попереднього діагнозу і визначенні тактики подальшого обстеження.

3. Для вдосконалення методу доцільно збільшити кількість обстежених хворих, врахувати додаткові показники, вибрані передусім так, щоб розрізнити затяжні пневмонії від раку й туберкульозу.

ЛИТЕРАТУРА

1. Афанасьева, Е. А. Алгоритмизация процессов ранней дифференциальной диагностики синдрома диабетической стопы: Автореф. дис.... канд. мед. наук. — Воронеж, 2008. — 20 с.
2. Батышин, И. З. Теория и практика нечетких гибридных систем [Текст] / Под ред. Н. Г. Ярушкиной. — Москва: Физматлит, 2006.
3. Насонов, С. В. Диагностика и дифференциальная диагностика заболеваний желудка с использованием высокоэффективной жидкостной хроматографии сыворотки крови: Автореф. дис.... канд. мед. наук. — Нижний Новгород, 2009. — 32 с.
4. Савула М. М. Дифференціальна діагностика захворювань легень і плеври [Текст] / За ред. М. М. Савули. — Тернопіль: Укрмедкнига, 2000. — 224 с.
5. Zimmermann, H.-J. Fuzzy Set Theory and Its Applications [Text] / H.-J. Zimmermann // 3rd ed. — Kluwer Academic Publishers, 1996. — 435 p.

ПРИМЕНЕНИЕ МОДИФИЦИРОВАННОГО КЛАСТЕРНОГО АНАЛИЗА ДАННЫХ СТАНДАРТНОГО МЕДИЦИНСКОГО ОБСЛЕДОВАНИЯ ДЛЯ ДИФФЕРЕНЦИАЛЬНОЙ ДИАГНОСТИКИ НЕКОТОРЫХ ЗАБОЛЕВАНИЙ ЛЕГКИХ

М. М. Савула, В. К. Паучок, Н. С. Кравченко

Резюме

Обследован 141 больной с патологией легких (рак легких — 35, туберкулез — 29, саркоидоз — 42, затяжные неспецифические пневмонии — 35). Предложен модифицированный метод кластерного анализа данных стандартного медицинского обследования с нелинейным преобразованием их координат. Получены качественные результаты, полезные для дифференциации саркоидоза, рака и туберкулеза. Для четкого выделения группы больных с затяжными пневмониями использованных данных недостаточно. Метод может быть полезным для установления предварительного диагноза на амбулаторном этапе и определения дальнейшей тактики врача.

MODIFIED CLUSTER ANALYSIS OF STANDARD MEDICAL EXAMINATION RESULTS IN DIFFERENTIAL DIAGNOSIS OF SOME PULMONARY DISEASES

М. М. Savula, V. K. Pauchok, N. S. Kravchenko

Summary

We examined 141 patients with lung diseases (cancer — 35, tuberculosis — 29, sarcoidosis — 42, pneumonia — 35). We developed a modified cluster analysis method of standard medical examination associated with non-linear data coordinates transformation. Obtained results are applicable for differentiation between sarcoidosis, cancer and tuberculosis. In the same time for significant determination of pneumonia these data are insufficient. The proposed approach is useful for preliminary diagnosis in outpatient department and further determination of diagnostics algorithm.