

УДК 519.217.2+616.006

А.М. Гупал, А.Л. Тарасов

**БАЙЄСІВСЬКИЙ ПІДХІД ПРИ РОЗПІЗНАВАННІ
ВТОРИННОГО ЕРИТРОЦИТОЗУ
ТА СПРАВЖНЬОЇ ПОЛІЦИТЕМІЇ**

Гупал Анатолій Михайлович

Інститут кібернетики імені В.М. Глушкова НАН України, м. Київ,
gupalanatol@gmail.com

Тарасов Андрій Леонтійович

Інститут кібернетики імені В.М. Глушкова НАН України, м. Київ,
freearcher@ukr.net

У статті вирішено завдання диференційної діагностики гематологічних захворювань — вторинного еритроцитозу та справжньої поліцитемії. Відомо, що на початковій стадії розвитку патологічних процесів клінічна картина та дані лабораторних досліджень при цих захворюваннях подібні. Аналіз лабораторних даних проведено з використанням байєсівських процедур розпізнавання. З 41 показника окремо виділено булеві та дискретні параметри. Вперше проведено аналіз одночасно трьох груп людей: група з вторинним еритроцитозом; група зі справжньою поліцитемією; група здорових людей. При проведенні аналізу булевих параметрів ефективність розпізнавання вторинного еритроцитозу та справжньої поліцитемії становить 100 %. Такий показник отримано завдяки тому, що в комбінації булевих параметрів C_n^k увійшли всі булеві параметри. При проведенні аналізу дискретних параметрів ефективність розпізнавання під час диференційної діагностики становить 95,55 %. При дослідженні трьох груп людей з використанням булевих параметрів ефективність розпізнавання становить 97,77 %, а з використанням дискретних параметрів — 86,66 %. Таким чином, побудувавши самонавчальну систему за допомогою байєсівських процедур розпізнавання, можна виконати швидку діагностику без широкого спектра аналізів пацієнта (4–7 параметрів з 41, що були проаналізовані). Найбільший внесок у досягнення високих результатів робить показник мутації JAK2, але, як показано в роботі, оцінку лабораторних досліджень можна проводити за іншими, більш доступними за вартістю показниками, такими як показники еритроцитів, тромбоцитів тощо. Даний підхід можна застосовувати і при діагностуванні інших захворювань, що потребують великої кількості аналізів.

Ключові слова: байєсівська процедура розпізнавання, булеві показники, дискретні показники, гематологічні захворювання, вторинний еритроцитоз, справжня поліцитемія.

Вступ

Еритроцитозом називається збільшення кількості еритроцитів у периферичній крові внаслідок їх інтенсивного утворення в кістковому мозку одночасно з підвищенням концентрації гемоглобіну, гематокриту та маси циркулюючих еритро-

© А.М. ГУПАЛ, А.Л. ТАРАСОВ, 2024

цитів. У роботі розглянуто два види еритроцитозів — вторинний еритроцитоз (ВЕ) та клональний еритроцитоз при справжній поліцитемії (СП). ВЕ та СП відносяться до гематологічних захворювань, пов'язаних з порушенням функцій крові. Як контрольна використана група здорових людей (ГЗЛ).

ВЕ — це стан, коли кількість еритроцитів в одиниці об'єму крові збільшується внаслідок активації еритропоезу та виходу надлишку еритроцитів з кісткового мозку в судинне русло.

СП — це злоякісний пухлинний процес у кровеносній системі. В крові з'являються надлишкові еритроцити і водночас збільшується кількість тромбоцитів та лейкоцитів, але меншою мірою.

Диференційна діагностика ВЕ, що виникає на фоні низки різноманітних патологічних станів, та клонального еритроцитозу при СП залишається складною задачею (особливо на початковій стадії патологічного процесу) через те, що клінічна картина та дані лабораторних показників при цих двох формах еритроцитозу можуть бути подібними [1, 2]. Так, близько 30 % хворих на СП при встановленні діагнозу не мають лейкоцитозу, тромбоцитозу та спленомегалії, які є критеріями діагностики цього захворювання, визначеними Всесвітньою організацією охорони здоров'я. З іншого боку, такі важливі діагностичні ознаки, як спленомегалія та лейкоцитоз, можуть бути присутні й у певної кількості пацієнтів з ВЕ.

У статті розглянуто байєсівський підхід до розпізнавання ВЕ, СП та ГЗЛ головним чином за стандартними клінічними дослідженнями крові. Основною метою при цьому підході є диференційна діагностика ВЕ та СП, а також розпізнавання одразу ВЕ, СП та ГЗЛ.

Розпізнавання на основі булевих параметрів

Булеві параметри — це параметри, які можуть приймати одне з двох можливих значень: *True* або *False*. У табл. 1 наведено 30 булевих показників дослідження крові та ефективності їх розпізнавання при диференційній діагностиці ВЕ та СП.

Таблиця 1

Булевий параметр	Розпізнавання, %		Булевий параметр	Розпізнавання, %	
	СП	ВЕ		СП	ВЕ
Мутація JAK2	94,44	98,33	Гіперплазія мегакаріоцитарна (МГКЦ) паростка	90,47	100
Еритропоетин, зниження	82,22	88,33	Трипаросткова гіперплазія	81,92	100
Гіперплазія мієлоїдного паростка	88,09	80,39	Нейтрофілоз	74,44	94,16
Мікроциркуляторні порушення	66,26	92	Спленомегалія по УЗД	65,55	85,83
Еритропоетин, норма	82,22	64,16	Лейкоцитоз	60	80
Холестерин, підвищення	95,55	59,16	Холестерин, норма	56,66	61,66
Тромбоцитоз	56,66	95,83	Плеоморфізм МГКЦ	55,95	100
Свербіж шкіри	53,33	85	Лактатдегідрогеназа (ЛДГ), норма	48,88	99,16
ЛДГ, підвищення	48,88	99,16	Холестерин, зниження	38,88	97,5
Феритин, норма	36,66	77,5	Феритин, зниження	32,22	92,5
Еритроцитоз	98,88	30,83	Учасник ліквідації наслідків аварії (УЛНА)	27,77	85
Гіперплазія еритроноу	100	19,6	Гемоглобін, підвищення	0	100
Резорбція кістки	39,28	0	Ретикуліновий фіброз	100	0
Феритин, підвищення	0	100	Гематокрит, підвищення	0	100
Еритропоетин, підвищення	0	24,16	Тромбоцити, зниження	0	100

У роботах [1, 2] виділено 15 булевих показників, ефективність розпізнавання яких вища за 50 %, та створено всі можливі комбінації цих незалежних ознак:

$C_{15}^2, C_{15}^3, \dots, C_{15}^{15}$. Водночас найвища ефективність розпізнавання досягнута в 98,04 % ВЕ та 100 % СП [2] для комбінації булевих параметрів: гіперплазія МГКЦ паростка, свербіж шкіри, мутація JAK2, плеоморфізм МГКЦ, трипаросткова гіперплазія та підвищення холестерину. Автори роботи не обмежились лише цими 15 параметрами, а поставили за мету досягти 100 %-го розпізнавання як для ВЕ, так і для СП шляхом побудови комбінацій булевих параметрів $C_{30}^2, C_{30}^3, \dots, C_{30}^8, \dots$. Розглянемо наведеної на рисунку алгоритм розпізнавання при диференційній діагностиці з використанням булевих параметрів (розподіл булевих показників при диференційній діагностиці ВЕ та СП).

Показники ВЕ		
Мутація JAK2	Холестерин, підвищення	Свербіж шкіри
30 показників <i>True</i>	25 показників <i>True</i>	20 показників <i>True</i>
90 показників <i>False</i>	95 показників <i>False</i>	100 показників <i>False</i>
Показники СП		
Мутація JAK2	Холестерин, підвищення	Свербіж шкіри
60 показників <i>True</i>	20 показників <i>True</i>	15 показників <i>True</i>
30 показників <i>False</i>	70 показників <i>False</i>	75 показників <i>False</i>

Процедура розпізнавання з незалежними ознаками базується на формулі Байєса

$$P(f = i | x_1, x_2, \dots, x_n) = \frac{P(x_1, x_2, \dots, x_n | f = i)P(f = i)}{P(x_1, x_2, \dots, x_n)}, \quad (1)$$

де $P(x_1, x_2, \dots, x_n | f = i)$ — добуток частот подій $x_1, x_2, \dots, x_n$ для діагнозів i ; $P(f = i)$ — частота появи діагнозу i у всій навчальній вибірці (кількість аналізів діагнозу i , поділена на загальну кількість аналізів); $P(x_1, x_2, \dots, x_n)$ — добуток частот подій $x_1, x_2, \dots, x_n$ у всій навчальній вибірці; $x_1, x_2, \dots, x_n$ — показники досліджень (на рисунку — мутація JAK2, підвищення холестерину та свербіж шкіри); i — діагнози, що розпізнаються (у формулах p — СП, e — ВЕ). Зазначимо, що при використанні цієї формули всі ймовірності в ній замінюються частотами з навчальної вибірки.

Відповідно до формули (1) при отриманні, наприклад, на вхід навчальної вибірки трьох показників (мутація JAK2 — *True*, підвищення холестерину — *False* та свербіж шкіри — *True*) маємо дві відповіді згідно з формулою (1) для СП та ВЕ. Стан дослідження відповідає максимальному з цих двох значень; при порівнянні відповідей знаменник у (1) скорочується та не використовується у розрахунках. Таким чином, отримуємо формулу

$$P(x_1, x_2, x_3)P(f = \{p, e\}) = \begin{cases} \frac{m_{1True_p} m_{2False_p} m_{3True_p} m_p}{m_p m_p m_p m_p + m_e}, \\ \frac{m_{1True_e} m_{2False_e} m_{3True_e} m_e}{m_e m_e m_e m_p + m_e}, \end{cases} \quad (2)$$

де m_{1True_p} , m_{2False_p} , m_{3True_p} — кількість показників *True* при мутації JAK2 (60 показників на рисунку), *False* — при підвищенні холестерину (70 показників на рисунку) та *True* — при свербіжі шкіри (15 показників на рисунку) відповідно при СП; m_{1True_e} , m_{2False_e} , m_{3True_e} — кількість показників *True* при мутації JAK2 (30 показників на рисунку), *False* — при підвищенні холестерину (95 показників на рисунку) та *True* — при свербіжі шкіри (20 показників на рисунку) відповідно при ВЕ; m_e , m_p — кількість аналізів для класів ВЕ та СП відповідно.

Ураховуючи, що верхня частина формули (2) порівнюється з нижньою, маємо

$$P(x_1, x_2, x_3)P(f = \{p, e\}) = \begin{cases} \frac{m_{1True_p} m_{2False_p} m_{3True_p}}{m_p^2}, \\ \frac{m_{1True_e} m_{2False_e} m_{3False_e}}{m_e^2}. \end{cases} \quad (3)$$

Згідно з рисунком та формулою (3) для СП отримуємо $60 \cdot 70 \cdot 15 / 90^2 = 7,78$ і для ВЕ — $30 \cdot 95 \cdot 20 / 120^2 = 3,96$. Вибираючи з двох значень максимальне, робимо висновок, що аналіз пацієнта стосується діагнозу СП. У табл. 2 наведено результати розпізнавання за комбінаціями булевих параметрів при диференційній діагностиці СП та ВЕ, а в табл. 3 — при розпізнаванні СП, ВЕ та ГЗЛ.

Таблиця 2

Комбінація параметрів	Розпізнавання, %	
	СП	ВЕ
Мутація JAK2	94,44	98,33
Мутація JAK2, трипаросткова гіперплазія	100	96,07
Мутація JAK2, гіперплазія МГКЦ паростка, еритроцитоз	98,8	100
Мутація JAK2, гіперплазія МГКЦ паростка, зниження еритропоєтину, норма еритропоєтину	100	100
Гіперплазія МГКЦ паростка ЛДГ, норма	95,23	100

Таблиця 3

Комбінація параметрів	Розпізнавання, %		
	СП	ВЕ	ГЗЛ
Мутація JAK2, підвищення гемоглобіну	94,44	81,66	100
Мутація JAK2, еритроцитоз, підвищення гемоглобіну	93,33	98,33	100
Мутація JAK2, еритроцитоз, підвищення ЛДГ, підвищення гемоглобіну	94,44	97,55	100
Мутація JAK2, норма еритропоєтину, зниження еритропоєтину, еритроцитоз, підвищення гемоглобіну, норма холестерину	97,77	99,16	100

З табл. 2 видно, що при диференційній діагностиці СП та ВЕ мета (100 %-ве розпізнавання) досягнута. Основний внесок при проведенні цього розпізнавання робить показник мутації JAK2 (розпізнавання тільки за цим показником не нижче 94,44 %). При мутації JAK2 йдеться про мутацію тірозінкінази, коли в 617-й позиції валін замінений фенілаланіном. Також досягнуте максимальне розпізнавання (98,8 %), коли з комбінації параметрів мутація JAK2 виключена. З табл. 3 видно, що максимальний показник при розпізнаванні одразу трьох станів не менший 97,77 %. У табл. 2, 3 показано, як зростає ефективність розпізнавання із зростанням кількості параметрів, що одразу входять у комбінацію параметрів (ці параметри є незалежними ознаками при байєсівському підході).

Розпізнавання СП та ВЕ для дискретного розподілу випадкових величин

Ще одна група показників має один показник у певному діапазоні значень. Особливість застосування байєсівських процедур до таких показників пов'язана з дискретизацією їхніх даних [3]. У табл. 4 наведено результати розпізнавання за такими показниками при диференційній діагностиці СП та ВЕ. У табл. 5 наведено диференційну діагностику СП та ВЕ, коли розглядаються комбінації дискретних параметрів з табл. 4.

З табл. 4 та 5 видно, що після збільшення кількості параметрів, які входять у комбінацію, підвищується ефективність розпізнавання. Так, за показниками лише тромбоцитів ефективність розпізнавання збільшується з 82,5 % до 95,55 %, коли в комбінаціях ураховується одразу п'ять параметрів.

Таблиця 4

Дискретний параметр	Розпізнавання, %		Дискретний параметр	Розпізнавання, %	
	СП	ВЕ		СП	ВЕ
Тромбоцити	83,33	82,5	Еритропоетин	82,22	85,83
Феритин	81,11	80,83	Сегментоядерні	76,66	75,83
Холестерин	80	74,16	ЛДГ	73,33	78,33
Еритроцити	71,11	73,33	Середній об'єм еритроцита (MCV)	72,22	70
Лейкоцити	68,88	69,16	Гематокрит	66,66	74,16
Гемоглобін	62,22	63,33	—	—	—

Таблиця 5

Комбінація параметрів	Розпізнавання, %	
	СП	ВЕ
Еритропоетин, сегментоядерні	88,88	89,16
Гематокрит, еритропоетин, тромбоцити	93,33	94,16
Еритропоетин, еритроцити, тромбоцити, феритин	94,44	95
MCV, гематокрит, еритропоетин, ЛДГ, тромбоцити	95,55	95,83

Як і при розпізнаванні СП, ВЕ та ГЗЛ з використанням комбінацій булевих параметрів, розглянемо результати розпізнавання зазначених станів людей за дискретними параметрами, наведеними в табл. 6.

Таблиця 6

Комбінація параметрів	Розпізнавання, %		
	СП	ВЕ	ГЗЛ
Еритроцити	67,77	73,33	100
Еритроцити, феритин	77,77	80,83	85,71
Еритроцити, сегментоядерні, феритин	82,22	82,5	85,71
Еритроцити, ЛДГ, тромбоцити, феритин	85,55	84,16	85,71
Гематокрит, еритроцити, ЛДГ, тромбоцити, феритин	85,55	85,83	85,71
MCV, гематокрит, еритроцити, сегментоядерні, тромбоцити, холестерин	85,55	86,66	96,42
MCV, гематокрит, еритроцити, ЛДГ, сегментоядерні, тромбоцити, холестерин	86,66	86,66	96,42

З табл. 6 видно, як із зростанням кількості дискретних параметрів, що одночасно враховуються при розпізнаванні, покращується ефективність розпізнавання. Так, за показниками лише еритроцитів ефективність зростає з 67,77 % до 86,66 %, коли в комбінацію параметрів одночасно входять сім параметрів (останній рядок табл. 6).

Висновок

При проведенні аналізу булевих параметрів розпізнавання СП та ВЕ досягнута 100 %-ва ефективність. Це стало можливим завдяки тому, що в комбінацію булевих параметрів C_n^k увійшли всі булеві параметри. Така висока ефективність розпізнавання досягнута при $n = 30$ та $k = 4$. У роботі [2] ефективність становить 98,04 % при $n = 15$ та $k = 6$. При аналізі дискретних параметрів розпізнавання ефективність диференційної діагностики становить 95,55 %. У роботі [2] вона становить 82,5 %.

Розпізнавання СП, ВЕ та ГЗЛ стало можливим завдяки тому, що було отримано нові дані для ГЗЛ. Ефективність розпізнавання за булевими показниками цих трьох станів не менша 97,77 %, а за дискретними показниками — 86,66 %.

Таким чином, побудувавши самонавчальну систему з використанням байєсівських процедур розпізнавання, можна виконати швидку діагностику без широкого спектра аналізів пацієнта. Даний підхід також можна застосовувати при діагностуванні інших захворювань, що потребують проведення великої кількості аналізів.

A. Gupal, A. Tarasov

BAYESIAN APPROACH IN RECOGNITION OF SECONDARY ERYTHROCYTOSIS AND POLYCYTHEMIA VERA

Anatolii Gupal

V.M. Glushkov Institute of Cybernetics of the NAS of Ukraine, Kyiv,
gupalanatol@gmail.com

Andrii Tarasov

V.M. Glushkov Institute of Cybernetics of the NAS of Ukraine, Kyiv,
freearcher@ukr.net

The task of differential diagnostics of hematological diseases such as secondary erythrocytosis and *True* polycythemia has been solved. It is known that at the initial stage of the development of pathological processes, the clinical picture and data of laboratory studies in these diseases are similar. Analysis of laboratory data was carried out using Bayesian recognition procedures. Boolean and discrete parameters are separately selected among 41 indicators. For the first time, an analysis of three conditions was carried out at once-secondary erythrocytosis, *True* polycythemia and a group of healthy people. When analyzing Boolean parameters, the efficiency of recognizing secondary erythrocytosis and *True* polycythemia was 100 %. It was possible to do this due to the fact that in this work, all Boolean parameters took part in the formation of combinations of Boolean parameters. When discrete parameters were analyzed, the recognition efficiency of differential diagnosis was 95,55 %. When recognizing *True* polycythemia, secondary erythrocytosis, and a group of healthy people at the expense of Boolean parameters, the efficiency of recognition was 97,77 %, and at the expense of discrete parameters 86,66 %. Thus, on the basis of a self-learning system due to Bayesian recognition procedures, it is possible to make a quick diagnosis without using a wide range of patient analyzes (4–7 parameters out of 41 that were analyzed). The greatest contribution to obtaining high results is provided by the JAK2 mutation indicator, but as shown in this work, the evaluation of laboratory studies can be carried out by other, more affordable indicators, such as indicators of erythrocytes, platelets, etc. This approach can be used to diagnose other diseases that require a large number of tests.

Keywords: Bayesian recognition procedure, Boolean indicators, discrete indicators, hematological diseases, secondary erythrocytosis, polycythemia vera.

ПОСИЛАННЯ

1. Вагис А.А., Гупал Н.А., Тарасов А.Л. Эффективные процедуры распознавания медицинских заболеваний. *Компьютерная математика*. 2014. № 2. С. 127–132.
2. Гупал А.М., Гупал М.А., Тарасов А.Л. Байєсівські процедури розпізнавання гематологічних захворювань. *Кібернетика і системний аналіз*. 2017. № 6. С. 118–124. <https://www.kibernetika.org/volumes/2017/numbers/06/articles/11/ArticleDetailsUA.html>
3. Тарасов А.Л. Аналіз показників поверхневого плазмонного резонансу з використанням байєсівських процедур розпізнавання з незалежними ознаками при гліомах, метастазах та менингіомах. *Кібернетика та комп'ютерні технології*. 2021. № 4. С. 35–42. DOI: <https://doi.org/10.34229/2707-451X.21.4.4>

Отримано 08.05.2024