УДК 519.217.2

А.М. ГУПАЛ, Н.А. ГУПАЛ, А.Л. ТАРАСОВ

БАЙЕСОВСКИЕ ПРОЦЕДУРЫ РАСПОЗНАВАНИЯ ГЕМАТОЛОГИЧЕСКИХ ЗАБОЛЕВАНИЙ

Аннотация. Обоснован перспективный компьютерный подход к распознаванию гематологических заболеваний. Вследствие высокой эффективности байесовских процедур путем перебора на компьютере подбираются такие комбинации показателей, которые обладают наиболее высоким качеством распознавания. Таким способом можно провести быструю диагностику, не выполняя ее в полном объеме.

Ключевые слова: байесовские процедуры распознавания, эритроцитозы, комбинация показателей.

ВВЕДЕНИЕ

Рассмотрим компьютерный подход распознавания гематологического заболевания на основе применения байесовских процедур с независимыми признаками при эритроцитозах и полицитемии. Эритроцитозом называется увеличение количества эритроцитов в периферической крови, как результат их интенсивного образования в костном мозге, в сочетании с повышением концентрации гемоглобина, показателей гематокрита и массы циркулирующих эритроцитов, превышающих нормальные возрастные и физиологические значения. При установлении диагноза сначала следует исключить истинную полицитемию (болезнь встречается часто у лиц пожилого возраста, но может быть у молодых людей и даже у детей). Установлена семейная предрасположенность к этому заболеванию, что говорит о его генетическом характере.

В настоящее время на основании молекулярно-генетических исследований создана определенная система классификации таких сложных и тяжелодиагностируемых заболеваний как эритроцитозы. Согласно этой системе эритроцитозы подразделяются на абсолютные, когда увеличиваются массы циркулирующих эритроцитов (МЦЭ), и относительные, обусловленные уменьшением объема циркулирующей плазмы (ОЦП). Абсолютные эритроцитозы, в свою очередь, подразделяются на первичные, которые обусловлены дефектами в эритроидных клетках-предшественниках, и вторичные. Вторичные эритроцитозы это состояния, характеризующиеся увеличением числа эритроцитов в единице объема крови в результате активации эритропоэза и выхода избытка эритроцитов из костного мозга в сосудистое русло. Как первичные, так и вторичные эритроцитозы могут быть врожденными и приобретенными.

В клинической практике сталкиваются с проблемами дифференциальной диагностики вторичного эритроцитоза — состоянием, при котором проявляются симптомы других болезней или патологических процессов. Устранение причин этих болезней или процессов приводит к ликвидации вторичных эритроцитозов без проведения специального лечения.

Истинная полицитемия (ИП) — это злокачественный опухолевый процесс системы крови. В крови появляется избыточное количество эритроцитов, однако при этом увеличивается количество тромбоцитов и лейкоцитов, но в меньшей степени. За счет увеличения числа эритроцитов повышается вязкость крови, возрастает масса циркулирующей крови. В результате происходит замедление кровотока в сосудах и образование тромбов, что приводит к нарушению кровоснабжения и

гипоксии. Истинная полицитемия относится к труднодиагностируемым заболеваниям. В 30–35% случаев у больных наблюдается только эритроцитоз, а такие характерные изменения, как лейкоцитоз и тромбоцитоз могут не проявляться на протяжении многих лет. Такой вариант ИП получил название чистый эритроцитоз. Только у трети больных может развиться типичная картина заболевания. Отметим, что ИП не является единственным заболеванием, при котором наблюдается увеличение красных показателей крови. Поэтому возникает необходимость в дифференциальной диагностике ИП и вторичных эритроцитозов.

Значительную роль в диагностике гематологических заболеваний сыграла расшифровка генетической информации человека и различных видов животных и растений. Это стало возможным при использовании современных компьютерных технологий и вычислительных методов.

Результаты клинических наблюдений и научные исследования показывают, что стандартные диагностические критерии ВОЗ для ИП не являются абсолютно удовлетворительными, а есть лишь предметом дальнейшего усовершенствования. Возможность выработки и усовершенствования эффективных диагностических критериев для распознавания данных гематологических заболеваний чрезвычайно важна особенно для Украины. Не следует забывать, что население Украины подверглось мощному влиянию негативних факторов аварии на Чернобыльской АЭС. Различные эритроцитозы могут быть спровоцированы действием радиоактивного излучения. Поэтому особо актуальным становится определение молекулярно-генетических критериев для диагностики подобных заболеваний.

РАСПОЗНАВАНИЕ ПОЛИЦИТЕМИИ И ВТОРИЧНОГО ЭРИТРОЦИТОЗА НА ОСНОВЕ БУЛЕВЫХ КОМБИНАЦИЙ

Рассмотрим задачу распознавания двух видов заболеваний: полицитемии и вторичного эритроцитоза. Диагностику проводили на основе анализа данных 41 показателя (анализа), позволяющих судить о наличии и степени заболевания. Отметим, что при дифференциальной диагностике полицитемии и эритроцитозов часто используют показатель мутации JAK2 (мутация тирозинкиназы JAK2 (янус-киназы)), где в позиции 617 валин заменен фенилаланином. Эта мутация наблюдается и при других гематологических заболеваниях, но при полицитемии она происходит наиболее часто.

Показатели для определения заболеваний образуют две группы. Первая группа содержит булевые показатели, а именно: гиперплазия эритрона; эритропоэтин повышение; зуд кожи; лактатдегидрогеназа (ЛДГ) повышение; микроциркуляторные нарушения; повышение гематокрита; резорбция кости; спленамогелия по УЗИ; тромбоциты снижение; ферритин повышение; холестерин норма; гиперплазия МГКЦ ростка; эритропоэтин снижение; мутация JAK2; гемоглобин повышение; ретикулиновый фиброз; триростковая гиперплазия; тромбоцитоз; ферритин снижение; холестерин повышение; гиперплазия миелоидного ростка; эритропоэтин норма; эритроцитоз; ЛДГ норма; лейкоцитоз; нейтрофилез; плеоморфизм МГКЦ; ликвидаторы аварии на ЧАЭС; ферритин норма; холестерин снижение. Булевы показатели принимают два значения: «да» или «нет». Количество таких показателей в обучающей выборке равно 30. Вторая группа содержит показатели, которые имеют диапазон значений. В данной обучающей выборке содержится 11 следующих показателей: эритропоэтин; ферритин; MCV; гематокрит; эритроциты; ЛДГ; лейкоциты; холестерин; гемоглобин; сегментоядерные; тромбоциты. Отметим, что на точность измерения некоторых показателей влияет ряд субъективных факторов, например ошибки специалистов при определении показателей. На фиксируемые показатели могут также влиять

принимаемые пациентами иные препараты, связанные с лечением других заболеваний, и т.д. Наиболее надежным показателем при установлении диагноза (эритроцитозов и полицитемии) является мутация JAK2, так как она не подвержена влиянию других внешних факторов [1].

Известно, что байесовская процедура распознавания дискретных объектов с независимыми признаками оптимальна [2]. Именно на этом основаны проводимые компьютерные расчеты. Как сказано выше, исследовались байесовские процедуры для 30 булевых показателей. Основная проблема заключается в выборе такой комбинации показателей, с помощью которой наилучшим образом можно проводить процесс распознавания указанных заболеваний. В обучающей выборке задействовано 90 больных с полицитемией и 120 — с эритроцитозами. Чтобы не проводить полного перебора всех комбинаций 30 показателей, подсчитывали качество распознавания каждого показателя в отдельности путем последовательного исключения больного из обучающей выборки.

Пусть в данном случае обучающая выборка V_0 содержит $m_0 = 90$ больных с полицитемией; выборка V_1 содержит $m_1 = 120$ больных с эритроцитозами; общее число больных составляет $m_2 = m_0 + m_1 = 210$. Байесовская процедура распознавания с независимыми признаками определяются по формуле

$$\xi(d, i) = \left(\frac{k(i)}{m_2} \right) \prod_{j=1}^{n} \left(\frac{k(d_j, i)}{m_i} \right), \quad i = 0, 1,$$

где $\frac{k(i)}{m_2}$ — априорная оценка состояния пациента, $\frac{k(d_j, i)}{m_i}$ — оценка вероятности $P(x_j | i)$, $x_j = d_j$, n — количество независимых признаков.

Для одного показателя ($n = 1$) процедура вычисления достаточно проста. Пусть анализируемый больной принадлежит выборке V_0 и $x_1 = d_1$; $k(0) = 89$, $k(1) = 120$, $m_2 = 209$; $k(d_1, i)$ — количество значений d_1 в обучающей выборке V_i. Поэтому

$$\xi(d,0) = \left(\frac{89}{209} \right) \left(\frac{k(d_1,0)}{89} \right), \ \xi(d,1) = \left(\frac{120}{209} \right) \left(\frac{k(d_1,1)}{120} \right).$$

Таким образом, состояние пациента определяется простым сравнением двух показателей: $k(d_1, 0)$ и $k(d_1, 1)$. Аналогичным образом подтверждается состояние пациента, которого исключают из выборки V_1.

В табл. 1 приведены 15 показателей, у которых эффективность распознавания заболеваний выше 50 %, т.е. число анализируемых показателей сократилось вдвое. Таким образом, количество комбинаций булевых показателей, которые необходимо проверить компьютерной системе, составляет $2^{15} - 1 = 32767$. Такое количество комбинаций позволяет получать результаты исследований на основе байесовской процедуры распознавания. На основе параметров из табл. 1 компьютерная система запускает байесовскую процедуру для всех возможных комбинаций булевых показателей.

В табл. 2 приведены некоторые комбинации показателей с наиболее высоким качеством распознавания двух заболеваний.

Если принимать во внимание такой исключительный показатель, как мутация JAK2, то согласно показателям в табл. 1 и 2 распознавание пациентов с полицитемией улучшилось с 94,44% до 100%, а для пациентов с эритроцитозами незначительно уменьшилось — с 98,33% до 98,04%. Таким образом, компьютерная система только по булевым показателям дает высокую эффективность распознавания.

Таблица 1. Булевы показатели с эффективностью распознавания выше 50 %

Название показателя	Показатель эффективности распознавания полицитемии, %	Показатель эффективности распознавания эритроцитозов, %
Мутация JAK2	94,4	98,3
Гиперплазия МГКЦ ростка	90,5	100
Эритропоэтин снижение	82,2	88,3
Триростковая гиперплазия	81,9	100
Гиперплазия миелоидного ростка	88,1	80,39
Нейтрофилез	74,4	94,3
Микроциркуляторные нарушения	66,3	92
Спленамогелия по УЗД	65,6	85,8
Эритропоэтин норма	82,2	64,2
Лейкоцитоз	60	80
Холестерин повышение	95,6	59,2
Тромбоцитоз	56,7	95,8
Холестерин норма	56,8	61,7
Плеоморфизм МГКЦ	55,9	55,9
Зуд кожи	53,3	85

Таблица 2. Результаты распознавания заболеваний по комбинациям булевых показателей

Номер комбинации	Комбинация показателей	Показатель эффективности распознавания полицитемии, %	Показатель эффективности распознавания эритроцитозов, %
1	Гиперплазия МГКЦ ростка Зуд кожи Мутация JAK2 Плеоморфизм МГКЦ Триростковая гиперплазия Холестерин повышение	100	98.04
2	Гиперплазия МГКЦ ростка Эритропоэтин снижение Нейтрофилез Триростковая гиперплазия	98,8	98,04
3	Гиперплазия МГКЦ ростка Лейкоцитоз Мутация JAK2 Триростковая гиперплазия Холестерин повышение	100	98.04
4	Гиперплазия МГКЦ ростка Мутация JAK2 Триростковая гиперплазия Холестерин повышение	100	98.04
5	Гиперплазия МГКЦ ростка Лейкоцитоз Мутация JAK2 Плеоморфизм МГКЦ Триростковая гиперплазия Холестерин повышение	100	98.04

Незначительное понижение эффективности распознавания пациентов с эритроцитозами объясняется тем, что выборки распознавания по мутации JAK2 (см. табл. 1) и по комбинациям показателей в табл. 2 отличаются. Поскольку обучающая выборка для групп показателей из табл. 2, как правило, меньше, чем для отдельных показателей, то, естественно, качество распознавания может снижаться.

Результаты вычислений в табл. 2 показывают, что компьютерная система путем перебора показателей выбирает наиболее информативные комбинации с числом показателей от четырех до шести. Таким образом, можно добиться быстрой диагностики, не проводя анализов всех 30 показателей.

РАСПОЗНАВАНИЕ ПОЛИЦИТЕМИИ И ВТОРИЧНОГО ЭРИТРОЦИТОЗА ДЛЯ ДИСКРЕТНОГО СЛУЧАЯ

Как упоминалось выше, вторая группа показателей имеет диапазон значений. Особенность применения байесовских процедур к таким показателям связана с дискретизацией их значений [3]. Результаты расчетов приведены в табл. 3. Таким образом, эффективность распознавания по показателям второй группы

Таблица 3. Результаты распознавания заболеваний для дискретного случая

Название показателя	Показатель эффективности распознавания полицитемии, %	Показатель эффективности распознавания эритроцитозов, %
Эритропоэтин	82,2	82,5
Тромбоциты	82,2	82,5
Ферритин	80	80,8
Сегментоядерные	76,6	75,8
Холестерин	80	74,1
ЛДГ	73,3	77,5
Эритроциты	72,2	73,3
MCV	72,2	70
Гемоглобин	62,2	63,3
Лейкоциты	58,8	80
Гематокрит	52,2	87,5

Таблица 4. Результаты распознавания заболеваний по комбинациям дискретных показателей

Номер комбинации	Комбинация показателей	Показатель эффективности до запуска распознавания по комбинации, %	Показатель эффективности распознавания комбинации, %
1	Тромбоциты Эритропоэтин Сегментоядерные Холестерин Эритроциты	82,5	90,0
2	Тромбоциты Холестерин Ферритин	82,5	88,8
3	Тромбоциты Эритропоэтин Холестарин	82,5	88,3
4	Эритропоэтин Сегментоядерные	82,2	84,4
5	Ферритин Сегментоядерные Холестерин	80,8	85,5

достигает приблизительно 80 %. В дальнейшем для распознавания диагнозов с полицитемией и эритроцитозами будут применены байесовские процедуры для всех возможных комбинаций показателей обучающей выборки.

Эффективность распознавания комбинации дискретных распределений для показателей тромбоцитов, эритропоэтина и сегментоядерных составила 91,1%, т.е. улучшение составило 8,6%. Такую же эффективность распознавания дают следующие комбинации: 1) тромбоциты, эритропоэтин, сегментоядерные, холестерин; 2) тромбоциты, эритропоэтин, ферритин; 3) тромбоциты, эритропоэтин, ферритин, эритроциты. В табл. 4 представлены другие комбинации дискретных распределений, которые дали меньшее, но все же весомое распознавание.

В табл. 5 приведены показатели эффективности распознавания байесовской процедуры для комбинаций булевых и дискретных величин. Отметим высокую эффективность байесовской процедуры распознавания для некоторых комбинаций, которая достигла 100% для двух классов показателей.

Таблица 5. Показатели эффективности распознавания заболеваний по комбинациям булевых и дискретных показателей

Номер комбинации	Комбинация показателей	Показатели эффективности до запуска распознавания по комбинации, %	Показатели эффективности распознавания комбинации, %
1	Мутация JAK2 Гиперплазия МГКЦ ростка Эритропоэтин Триростковая гиперплазия Гиперплазия миелоидного ростка Нейтрофилез Эритроциты	93,98	100
2	Мутация JAK2 Гиперплазия МГКЦ ростка Триростковая гиперплазия Эритроциты	93,98	100
3	Мутация JAK2 Гиперплазия МГКЦ ростка Эритропоэтин Эритроциты	94,05	100

ЗАКЛЮЧЕНИЕ

В настоящей статье проводилось распознавание гематологических заболеваний при эритроцитозах с использованием байесовских процедур. Количество показателей или исследований, которые фиксировались у пациентов, в данной работе составляют 41 показатель. Из них выбираются такие, которые каждый в отдельности дает распознавание выше 50%. Наиболее значимым из этих показателей является мутация JAK-2, исходя из которой эффективность распознавания при эритроцитозах составляет приблизительно 95%. Были получены три комбинации показателей, что составляет 100% распознавания гематологических заболеваний. В каждой комбинации основной вклад дает мутация JAK-2. Отметим, что необязательно в комбинациях должна присутствовать эта

мутация. На основе других комбинаций показателей также можно осуществлять распознавание с высокой эффективностью.

Таким образом, на основе самообучающейся системы с использованием байесовских процедур распознавания удается проводить быструю диагностику, не задействуя широкого спектра анализов для пациента. Данный подход можно применять и при диагностировании других заболеваний, требующих большого количества анализов.

СПИСОК ЛИТЕРАТУРЫ

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., Сергиенко И.В. Оптимальные процедуры распознавания. Киев: Наук. думка, 2008. 232 с.
3. Гупал А.М., Сергиенко И.В. Симметрия в ДНК. Методы распознавания дискретных последовательностей. Киев: Наук. думка, 2016. 228 с.

Надійшла до редакції 19.05.2017

А.М. Гупал, М.А. Гупал, А.Л. Тарасов
БАЄСІВСЬКІ ПРОЦЕДУРИ РОЗПІЗНАВАННЯ ГЕМАТОЛОГІЧНИХ ЗАХВОРЮВАНЬ

Анотація. Обгрунтовано перспективний комп'ютерний підхід до розпізнавання гематологічних захворювань. Внаслідок високої ефективності баєсівських процедур шляхом перебору на комп'ютері підбираються такі комбінації показників, які мають найвищу якість розпізнавання. У такий спосіб можна отримати швидку діагностику, не виконуючи її в повному обсязі.

Ключові слова: баєсівські процедури розпізнавання, еритроцитози, комбінація показників.

A.M. Gupal, M.A. Gupal, A.L. Tarasov
BAYESIAN PROCEDURES OF RECOGNITION OF HEMATOLOGY DISEASES

Abstract. A promising computer approach to recognition of hematology diseases is substantiated. Due to fast operation of Bayesian procedures, computer search is used to find combinations of indicators that have the highest recognition quality. Such method allows conducting fast diagnostics without performing it completely.

Keywords: Bayesian recognition procedures, erythrocytosis, combination of indicators.

Гупал Анатолий Михайлович,
чл.-кор. НАН Украины, доктор физ.-мат. наук, профессор, заведующий отделом Института кибернетики им. В.М. Глушкова НАН Украины, Киев, e-mail: gupal_anatol@mail.ru.

Гупал Никита Анатольевич,
научный сотрудник Института кибернетики им. В.М. Глушкова НАН Украины, Киев,
e-mail: nikita_gupal@yahoo.com.

Тарасов Андрей Леонтьевич,
кандидат техн. наук, научный сотрудник Института кибернетики им. В.М. Глушкова НАН Украины, Киев,
e-mail: freearcher@ukr.net.