УДК 004.9:614.2
DOI: 10.15587/2706-5448.2020.210671

ПРОЕКТУВАННЯ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ КЛАСИФІКАЦІЇ НА ОСНОВІ МЕДИЧНИХ ДАНИХ

Мулеса О. Ю., Снитюк В. Є., Тромбола М. І., Іваськевич В. З.

Процеси прийняття рішень, пов'язані з віднесенням особи до групи ризику виникнення захворювань, супроводжуються необхідністю аналізу великих обсягів медико-соціальних даних. При цьому, кваліфікований лікар має оперувати як особистими даними пацієнта, так і релевантними протоколами лікування та інструкціями. Раннє прогнозування ризиків виникнення захворювань дозволяють медичним працівникам здійснювати планування, розробляти систему превентивних заходів тощо. Тому об'єктом дослідження є підтримка та інформаційно-аналітичний супровід процесів прийняття рішень щодо раннього прогнозування ризиків виникнення захворювань у осіб на основі медичних даних. Такий супровід необхідний для аналізу досвіду медичного працівника, який зафіксований у вигляді статистичних даних. Одним з найбільш проблемних місць на етапі проектування та впровадження релевантної інформаційної технології є збір та аналіз статистичних даних щодо досліджуваної проблеми.

Дослідження було виконано відповідно до методології системного підходу. Всі етапи проектування інформаційної технології прогнозування на основі медичних даних відповідають етапам системного підходу: систематизації, формалізації, цілеорієнтації. В основі розробленої технології лежить метод класифікації на основі послідовного аналізу Вальда.

В результаті дослідження було:

– побудовано математичну модель задачі прогнозування ризиків виникнення захворювання як задачі класифікації;

– розроблено функціональну схему інформаційно-аналітичної системи для розв'язання задачі класифікації на основі медичних даних. Аналітичне ядро інформаційно-аналітичної системи утворюють алгоритми статистичної обробки даних, а також метод класифікації на основі послідовного аналізу Вальда;

– виконано експериментальну верифікацію розробленої технології для задачі прогнозування виникнення зубощелепних аномалій у дітей. На основі наявних статистичних даних була побудована диференціально-прогностична таблиця. Виконано всі обчислення. На прикладах продемонстровано ефективність розробленої технології.

Розроблена інформаційна технологія може використовуватися медичними працівниками у процесі здійснення раннього прогнозування ризиків виникнення хвороб.

Ключові слова: *медико-соціальні дані, медична статистика, групи ризику, медичні послуги, інформаційно-аналітичний супровід.*

1. Вступ

Процеси прийняття рішень, пов’язані з віднесенням особи до групи ризику виникнення захворювань супроводжуються необхідністю аналізу великих обсягів медико-соціальних даних. При цьому, кваліфікований лікар має оперувати як особистими даними пацієнта, так і релевантними протоколами лікування та інструкціями. Проте, на практиці, додатковим важливим джерелом даних є власний досвід лікаря, який фіксується у вигляді медичної статистики. Спираючись на досвід попередніх випадків виникнення у різних осіб захворювання, кваліфікований лікар може виконати ранній прогноз та віднести особу до групи ризику виникнення захворювання. Виконання раннього прогнозу дозволяє медичному працівникові запланувати та реалізувати ряд превентивних заходів, надати особі кваліфіковану допомогу та рекомендації щодо подальших дій.

Враховуючи те, що описані процеси прийняття рішень вимагають від лікаря оперативного та точного опрацювання даних різної природи, актуальною є розробка та впровадження релевантних інформаційних технологій віднесення осіб до груп ризику.

2. Об’єкт дослідження та його технологічний аудит

Об’єктом дослідження є підтримка та інформаційно-аналітичний супровід процесів прийняття рішень щодо раннього прогнозування ризиків виникнення захворювань у осіб на основі медичних даних. Такий супровід необхідний для аналізу досвіду медичного працівника, який зафіксований у вигляді статистичних даних. На основі виконаного аналізу можливим є прийняття рішення щодо віднесення особи до відповідної групи ризику з метою реалізації щодо неї подальших заходів медичного характеру.

На етапі проектування та впровадження релевантної інформаційної технології важливими є збір та аналіз статистичних даних щодо досліджуваної проблеми.

3. Мета та задачі дослідження

Метою дослідження є виконання аналізу етапів процесу проектування інформаційної технології віднесення особи до групи ризику виникнення захворювання.

Для досягнення поставленої мети необхідно виконати такі завдання:

1. Здійснити систематизацію проблем і завдань, які виникають в процесі проектування релевантної інформаційної технології.
2. Побудувати математичну модель задачі прогнозування ризиків виникнення захворювання як задачі класифікації.
3. Розробити функціональну схему інформаційно-аналітичної системи для розв’язання задачі класифікації на основі медичних даних.
4. Виконати експериментальну верифікацію розробленої технології для задачі прогнозування виникнення зубощелепних аномалій у дітей.

4. Дослідження існуючих шляхів вирішення проблеми

Проблемам прогнозування в медицині присвячено ряд наукових досліджень. В [1] наведено дослідження, в якому виконано прогноз основних

показників, пов’язаних з виникненням та лікуванням ниркової недостатності. Прогнозування виконано на основі динамічних рядів різними прогнозними моделями. Робота [2] присвячена проблемі застосування моделей та методів великих даних (Big Data) для аналізу медичних даних в ході розв’язання задач прогнозування та класифікації. Тут окремо розглянуто питання використання хмарних додатків для обробки та аналізу медичних даних. В [3, 4] міститься детальний аналіз проблеми застосування методів прогнозування динамічних рядів для моделювання та прогнозування потреб в невідкладній медичній допомозі в регіоні. Для прогнозування було використано моделі авторегресії з різними параметрами. Проблемі прогнозування майбутніх потреб в медичних послугах присвячено [5]. Прогнозування в роботі здійснюється на основі часових рядів, які характеризують основні демографічні показники регіону, а також з використанням імітаційних моделей. Система керування медичними даними з використанням механізмів прогнозування медичних подій приведена в [6]. Розроблена система працює як із статистичними, так і динамічними даними. Прогнозування здійснюється для різних когорт. В основі системи лежать моделі та методи машинного навчання. Порівняльний аналіз сучасних методів прогнозування медичних динамічних рядів наведено в [7, 8].

Аналітичний огляд сучасних наукових публікацій, присвячених прогнозуванню на основі медичних даних, показав, що їх переважна більшість містить моделі та методи прогнозування на основі часових рядів. Задачі прогнозування ризиків виникнення в майбутньому таких станів, які потребуватимуть медичного супроводу, для конкретних клінічних випадків розв’язуються рідко. Для розв’язання такого типу задач успішно застосовують патометричні алгоритми, які базуються на аналізі статистичних даних та розробленні чітких правил вироблення прогнозу [9, 10]. В основі таких алгоритмів, як правило лежить теорема Байєса.

Таким чином, доцільною є розробка інформаційних технологій прогнозування ризиків виникнення захворювань на основі ймовірнісних методів.

5. Методи дослідження

Дослідження було виконано відповідно до методології системного підходу. Всі етапи проектування інформаційної технології прогнозування на основі медичних даних відповідають етапам системного підходу.

В основі розробленої технології лежить метод класифікації на основі послідовного аналізу Вальда.

6. Результати дослідження

Інформаційна технологія віднесення особи до групи ризику виникнення захворювань була спроектована відповідно до етапів системного підходу – систематизації, формалізації, цілеорієнтації [9].

На етапі систематизації проблем та завдань, для реалізації яких призначена інформаційна технологія було проаналізовано проблему раннього прогнозування ризику виникнення захворювання на основі медичних даних. Вхідними даними при виконанні такого прогнозу є ретроспективні дані про

випадки виникнення захворювання в минулому у різних осіб. Важливою проблемою при цьому є аналіз репрезентативності наявної вибірки даних.

Другим проблемним питанням є вибір ознак, за якими можливо виконати раннє прогнозування. Випадок, при якому раннє прогнозування ризику виникнення захворювань можливо зробити тільки на основі результатів медичних обстежень особи є тривіальним і, як правило, не потребує додаткових досліджень. В загальному ж випадку, медичному працівнику необхідно аналізувати результати анамнезу особи, в які входять її медико-соціальні дані.

Таким чином, інформаційна технологія, призначенням якої є інформаційно-аналітичний супровід процесів прийняття рішень щодо прогнозування ризику виникнення захворювання у особи, має включати у себе моделі та методи аналізу статистичних даних.

Формально задачу прогнозування ризику виникнення захворювання у особи можна представити у вигляді задачі класифікації таким чином [10]: нехай маємо множину об'єктів:

$$O = \{O_1, O_2, ..., O_n\},$$

для кожного з яких відомі значення за кожним критерієм з множини:

$$K = \{K_1, K_2, ..., K_m\}.$$

Тобто, задана множина векторів:

$$W = \{w_i = (w_{i1}, w_{i2}, ..., w_{im}), i = \overline{1,n}\},$$

де w_{ij} – значення j-ого критерію для i-го об'єкту. Кожен об'єкт належить до одного з двох заданих класів A та B, при чому перший клас відповідає наявності ризику виникнення захворювання, а другий – відсутності такого ризику. Необхідно задати правило, за яким для деякого об'єкта O', який характеризується вектором $w' = (w'_1, w'_2, ..., w'_m)$ відповідних критеріїв з множини K, на основі даних про об'єкти з множини O, можна буде прийняти рішення про його віднесення до одного з класів A або B.

Для розв'язання поставленої задачі було обрано метод класифікації на основі послідовного аналізу Вальда [10]. Для застосування даного методу, на початковому етапі необхідно виконати попередню обробку вхідних даних [10, 11]:

– обчислити діагностичні коефіцієнти критеріїв множини K, які базуються на понятті умовної імовірності;

– встановити інформативність кожного критерію на основі міри Кульбака;

– визначити рівні надійності прийнятих рішень, які пов'язані з показниками помилок першого та другого року: під помилкою першого роду розуміють помилкове віднесення об'єкта до класу B, відповідно помилка другого роду описує помилкове віднесення об'єкта до класу A.

Власне, метод класифікації на основі послідовного аналізу Вальда полягає в тому, що для заданого об’єкта, на основі навчальної вибірки, обчислюються суми, які характеризують його близькість до класів.

Процедури, які реалізують етапи попередньої обробки даних та метод класифікації на основі послідовного аналізу Вальда утворюють аналітичне ядро інформаційної технології. В результаті роботи процедур, для кожного об’єкта можливим є отримання одного з таких рішень:

– об’єкт відноситься до класу A з мірою належності μ_A;

– об’єкт відноситься до класу B з мірою належності μ_B;

– для прийняття рішення при заданих умовах недостатньо інформації.

На основі згаданих моделей та методів була розроблена інформаційно-аналітична система класифікації на основі медичних даних. Функціональна схема системи наведена на рис. 1.

Рис. 1. Функціональна схема інформаційно-аналітичної системи класифікації на основі медичних даних

Для експериментальної верифікації спроектованої інформаційної технології (ІТ) була розв’язана задача прогнозування ризиків виникнення зубощелепних аномалій у дітей та юнацтва.

На етапі була створена база даних про 105 дітей віком 7–9 років. Відповідно до двох прогнозованих станів, на основі бази даних було сформовано дві навчальні вибірки: 60 дітей, у яких було виявлено зубощелепні аномалії у 13 років, і 45 дітей, у яких не було зубощелепних аномалій.

Задача полягає у тому, щоб для конкретної дитини віком 7–9 років на основі аналізу факторів ризику вибрати одне з двох прогностичних рішень: перше – у дитини високий ризик виникнення зубощелепних аномалій (стан А), друге – низький ризик виникнення аномалій (стан В). При цьому, значення порогів для A=6.4, для B=–6.4.

Початкові дані наведені в табл. 1.

Таблиця 1

Початкові дані для задачі класифікації

Показник	Значення показника	Група обстежених	
		Основна (n=60)	Контроль (n=45)
1. Спадковість	0–1	15	35
	2–6	45	10
2. Шкідливі звички	Так	44	12
	Ні	16	33
3. Травми	Так	20	11
	Ні	40	34
4. Патологічна постава	Так	32	23
	Ні	28	22
5. Ранній каріє с і раннє видалення зубів	Так	40	12
	Ні	20	33
6. Неправильне штучне вигодовування, довготривале смоктання пустишки, вживання м’якої їжі	Так	34	10
	Ні	26	35
7. Неправильна закладка зачатків зубів	Так	3	2
	Ні	57	43
8. ЛОР-патології	Так	25	8
	Ні	35	37
9. Системні та хромосомні захворювання організму	Так	1	1
	Ні	59	44
10. Екологічно-гігієнічні фактори	Так	25	10
	Ні	35	35
11. Патології перебігу вагітності матері	Так	30	10
	Ні	30	35
12. Тривала заміна тимчасових зубів постійними	>16	10	25
	0–16	50	20

Відповідно до функціональної схеми ІТ, початкові дані були оброблені методом нечіткої класифікації на основі послідовного аналізу Вальда, в результаті чого були отримані такі діагностичні коефіцієнти та рівні інформативності для показників (табл. 2).

Таблиця 2

Диференціально-прогностична таблиця

Показник	Значення показника	Діагностичний коефіцієнт	Інформативність
1. Спадковість	0–1	–4,93	2,69
	2–6	5,28	
2. Шкідливі звички	Так	4,39	2,05
	Ні	–4,39	
3. Травми	Так	1,35	0,08
	Ні	–0,54	
4. Патологічна постава	Так	0,18	0,004
	Ні	–0,20	
5. Ранній карієс і раннє видалення зубів	Так	3,98	1,48
	Ні	–3,42	
6. Неправильне штучне вигодовування, довготривале смоктання пустишки, вживання м’якої їжі	Так	4,07	1,18
	Ні	–2,54	
7. Неправильна закладка зачатків зубів	Так	0,51	0,001
	Ні	–0,03	
8. ЛОР-патології	Так	3,70	0,62
	Ні	–1,49	
9. Системні та хромосомні захворювання організму	Так	–1,25	0,004
	Ні	0,02	
10. Екологічно-гігієнічні фактори	Так	2,73	0,39
	Ні	–1,25	
11. Патології перебігу вагітності матері	Так	3,52	0,76
	Ні	–1,92	
12. Тривала заміна тимчасових зубів постійними	>16	–5,23	1,55
	0–16	2,73	

Так як, відповідно до методу нечіткої класифікації, приймати до уваги потрібно тільки ті показники, інформативність яких перевищує 0,5, то класифікація проводитиметься за такими показниками:

– спадковість;
– шкідливі звички;
– ранній карієс і раннє видалення зубів;
– неправильне штучне вигодовування;
– ЛОР-патології;
– екологічно-гігієнічні фактори;

– патології перебігу вагітності у матері;

– тривала заміна тимчасових зубів.

Для ілюстрації роботи алгоритму розглянемо приклад даних дитини віком 10 років, у якої пізніше, у 13 років, були виявлені зубощелепні аномалії.

Приклад 1. Дитина 10 років. Застосуємо алгоритм прогнозування, розглядаючи значення показників дитини в порядку зменшення їх інформативності. Показники дитини та результати роботи алгоритму наведені в табл. 3.

Таблиця 3

Результати роботи прогностичного алгоритму (Приклад 1)

Крок алгоритму	Показник	Значення	Сума діагностичних коефіцієнтів	Аналіз
1	Спадковість	4	S=5,28	–6,5<5,28<6,5
2	Шкідливі звички	Ні	S=5,28–4,39=0,89	–6,5<0,89<6,5
3	Тривала заміна тимчасових зубів постійними	12	S=0,89+2,73=3,62	–6,5<3,62<6,5
4	Ранній карієс і раннє видалення зубів	Так	S=3,62+3,98=7,6	7,6>6,5

Як видно з табл. 3, алгоритм зупинив свою роботу на четвертому кроці, коли сума діагностичних коефіцієнтів перевищила поріг. Таким чином, було прийнято рішення віднести пацієнта до групи А високого ризику виникнення зубощелепних аномалій.

Приклад 2. Дитина 9 років, у якої в підлітковому віці не було виявлено зубощелепних аномалій. Результати роботи прогностичного алгоритму наведені в табл. 4.

Таблиця 4

Результати роботи прогностичного алгоритму (Приклад 2)

Крок алгоритму	Показник	Значення	Сума діагностичних коефіцієнтів	Аналіз
1	Спадковість	3	S=5,28	–6,5<5,28<6,5
2	Шкідливі звички	Ні	S=5,28–4,39=0,89	–6,5<0,89<6,5
3	Тривала заміна тимчасових зубів постійними	17	S=0,89–5,23=–4,34	–6,5<–4,34<6,5
4	Ранній карієс і раннє видалення зубів	Ні	S=–4,34–3,42=–7,76	–7,76<–6,5

В даному прикладі (табл. 4) видно, що алгоритм зупинився на етапі, коли сума діагностичних коефіцієнтів стала меншою за нижній поріг. В результаті

цього було прийняте рішення – віднести пацієнта у групу В низького ризику виникнення зубощелепних аномалій.

Таким чином, спроектована інформаційна технологія може успішно використовуватися лікарями-стоматологами на ранніх етапах для прогнозування можливості ризиків виникнення аномалій в майбутньому.

7. SWOT-аналіз результатів дослідження

Strengths. Впровадження розробленої інформаційної технології дозволить на ранніх етапах здійснювати прогнозування ризиків виникнення захворювання. Виконання точного та вчасного прогнозу дозволить медичному працівнику розробити комплекс превентивних заходів щодо кожного конкретного клінічного випадку і тим самим, зменшити можливі ризики виникнення небезпечних станів у пацієнта.

Weaknesses. Особливістю розробленої технології є те, що для правильності виконаного медичного прогнозу медичний працівник має сформувати репрезентативні навчальну та контрольні вибірки на основі власного досвіду або доступних статистичних даних.

Opportunities. В процесі розвитку спроектованої інформаційної технології доцільним є впровадження елементів теорії нечітких множин в опис основних показників, за якими здійснюється прогнозування. Такий підхід дозволить приймати ефективні рішення у випадку, коли показники приймають значення, близькі до граничних.

Threats. Застосування розробленої інформаційної технології в різних медичних сферах потребує виконання попереднього аналізу проблеми для формування множини факторів ризику виникнення захворювання, а також інтервалів значень, які ці фактори можуть приймати.

8. Висновки

1. Здійснено систематизацію проблем і завдань, які виникали в процесі проектування інформаційної технології прогнозування ризиків виникнення захворювання. Відзначено, що вхідними даними в процесі функціонування технології є ретроспективні дані про випадки виникнення захворювання в минулому у різних осіб.

2. Побудовано математичну модель задачі прогнозування ризиків виникнення захворювання як задачі класифікації. Класифікація здійснюється шляхом віднесення об’єкта до одного з двох класів: класу з наявним ризиком виникнення захворювання та клас без такого ризику.

3. Розроблено функціональну схему інформаційно-аналітичної системи для розв’язання задачі класифікації на основі медичних даних. Аналітичне ядро інформаційно-аналітичної системи утворюють алгоритми статистичної обробки даних, а також метод класифікації на основі послідовного аналізу Вальда.

4. Виконано експериментальну верифікацію розробленої технології для задачі прогнозування виникнення зубощелепних аномалій у дітей. В ході верифікації було сформовано навчальну та контрольну вибірки та відібрано

фактори ризику (показники) виникнення вказаних аномалій. На двох прикладах продемонстровано ефективність розробленої інформаційної технології.

Література

1. Sun, L., Zou, L.-X., Han, Y.-C., Huang, H.-M., Tan, Z.-M., Gao, M. et. al. (2016). Forecast of the incidence, prevalence and burden of end-stage renal disease in Nanjing, China to the Year 2025. *BMC Nephrology, 17 (1).* doi: http://doi.org/10.1186/s12882-016-0269-8
2. Li, J.-S., Zhang, Y.-F., Tian, Y. (2016). Medical big data analysis in hospital information system. *Big data on real-world applications*, 65. doi: http://doi.org/10.5772/63754
3. Juang, W.-C., Huang, S.-J., Huang, F.-D., Cheng, P.-W., Wann, S.-R. (2017). Application of time series analysis in modelling and forecasting emergency department visits in a medical centre in Southern Taiwan. *BMJ Open, 7 (11),* e018628. doi: http://doi.org/10.1136/bmjopen-2017-018628
4. Steins, K., Matinrad, N., Granberg, T. (2019). Forecasting the Demand for Emergency Medical Services. *Proceedings of the 52nd Hawaii International Conference on System Sciences.* doi: http://doi.org/10.24251/hicss.2019.225
5. Lopes, M. A., Almeida, Á. S., Almada-Lobo, B. (2016). Forecasting the medical workforce: a stochastic agent-based simulation approach. *Health Care Management Science, 21 (1),* 52–75. doi: http://doi.org/10.1007/s10729-016-9379-x
6. Park, Y., Ho, J., Vishwanath, S. (2016). *U.S. Patent Application No. 15/092,738.*
7. Amor, L. B., Lahyani, I., Jmaiel, M. (2016). Recursive and Rolling Windows for Medical Time Series Forecasting: A Comparative Study. *2016 IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES),* 106–113. doi: http://doi.org/10.1109/cse-euc-dcabes.2016.169
8. Kristianto, R. P., Utami, E. (2017). Optimization the parameter of forecasting algorithm by using the genetical algorithm toward the information systems of geography for predicting the patient of dengue fever in district of sragen, Indonesia. *2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE),* 45–50. doi: http://doi.org/10.1109/icitisee.2017.8285548
9. Tymchenko, A. A. (2005). Systemnyi pidkhid do naukovoho doslidzhennia (orhanizatsiino-metodychni aspekty). *Visnyk ChDTU, 1,* 191–197.
10. Mulesa, O. Yu., Snytiuk, V. Ye., Herzanych, S. O. (2020). A fuzzy classification method based on the sequential wald analysis. *Automation of technological and business processes, 11 (4),* 35–42. doi: http://doi.org/10.15673/atbp.v11i4.1597
11. Herzanych, S. O., Mulesa, O. Yu. (2018). Alhorytm prohnozuvannia nevynoshuvannia vahitnosti v umovakh pryrodnoho yodnoho defitsytu. *Zdorove zhenshchyni, 8 (134),* 48–51.