

КАФЕДРА ІНФОРМАЦІЙНИХ СИСТЕМ ТА МЕРЕЖ ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ ВИКОРИСТАННЯ ДЕРЕВ ПРИЙНЯТТЯ РІШЕНЬ ДЛЯ ПРОГНОЗУВАННЯ ДІАГНОЗУ В МЕДИЦИНІ*

© Нікольський Ю.В., Щербина Ю.М., 2004

Проаналізовано результати застосування дерев рішень, побудованих за алгоритмом С 4.5, з точки зору оцінки їх ефективності для прогнозування діагнозу певного кардіологічного захворювання. Досліджено вплив способу вибору даних на якість прогнозу.

The analysis the results of application the decisions trees for constructing an algorithm С 4.5 for forecasting the diagnosis of some cardiological disease is carried out. The influence of a way to choice the data on quality of the forecasting is researched.

Постановка проблеми в загальному вигляді

Постановка діагнозу в медицині базується на загальній процедурі прийняття рішень, яка використовує знання лікаря, які сформовані в процесі його діяльності, інформацію про хворого та вміння використати знання та інформацію для перевірки гіпотези щодо прогнозованого діагнозу. Моделювання процедур прийняття рішень, основаних на використанні дедуктивного підходу для перевірки гіпотез, вимагає побудови баз знань, які якнайширше використовують знання фахівців з конкретної медичної проблемної сфери. Традиційний підхід до створення інтелектуальних інформаційних систем прийняття рішень в медицині вимагає формування баз знань як результату взаємодії інженера з експертом. Такий підхід до створення традиційних експертних систем має свої недоліки [1], пов'язані із специфікою отримання знань від експерта в процесі конструювання бази знань. Сучасні підходи, які основані на індуктивному підході, використовують технології машинного навчання [2–3]. Такі підходи базуються на формуванні та аналізі дуже великих баз даних [4], в яких накопичується вся можлива інформація про предметну галузь.

У практичних задачах, результати дослідження яких наводяться у цій статті, приймається рішення відносно досліджуваних об'єктів шляхом віднесення їх до класів або груп наданням відповідних ознак цим об'єктам. Якби такі ознаки були невідомими, то було б необхідно розбити множину об'єктів на підмножини подібних у деякому сенсі об'єктів та визначити ознаку кожної підмножини. Такі задачі називають *задачами кластеризації*. Для досліджуваної проблеми ці ознаки відомі наперед, і задача є *задачею класифікації*.

Для об'єктів, щодо яких рішення вже прийняте, взагалі кажучи, немає інформації про правила отримання цими об'єктами класифікаційних ознак. Тому задача прийняття рішень щодо нових об'єктів полягає або у знаходженні правил, за якими такі ознаки визначались, або у пошуку спільних властивостей цих об'єктів. Як перший, так і другий клас задач є задачами машинного навчання в тому сенсі, що правила визначення ознак треба шукати на основі інформації про об'єкти, яким такі ознаки вже надані або у пошуку закономірностей на множині об'єктів, для яких ознаки прийняття рішень не визначені. Знайдені правила надалі будуть використані для прийняття рішення

* Дослідження, результати яких подано у статті, частково підтримані грантом від бюро з питань освіти та культури (ЕСА) Держдепартаменту США. ЕСА не несе відповідальності за погляди, що тут висловлені.

щодо нових об'єктів, яким потрібно буде надавати такі ознаки. Відповідні алгоритми пошуку правил знаходження рішень у цьому формулюванні називають *алгоритмами* машинного навчання [2–3] та реалізують у вигляді спеціальних процедур, що побудовані за індуктивним принципом: на основі наявних прикладів шукають приховані закономірності, які використовують для подальшого розв'язування задач прийняття рішень у нових прикладах, для яких значення ознак є невідомим. Тут термін „приклад” використано в інтуїтивному розумінні, а точне означення буде наведено нижче.

Методи класифікації називають методами *навчання з учителем (supervised learning)*, відповідні ознаки – *атрибутами прийняття рішень*, а таблиці характеристик об'єктів – *таблицями прийняття рішень*. Методи кластеризації називають методами *навчання без учителя (unsupervised learning)*: у них наперед невідома кількість класів, на які буде розбита множина об'єктів та відповідні мітки цих класів. До основних методів класифікації даних належать методи, що ґрунтуються на побудові дерев рішень, *k*–найближчих сусідів, міркування на підставі попередніх випадків, наближених множин (rough sets) тощо. Задачі кластеризації розв'язують методами кластерного аналізу або спеціальними класами нейромереж.

У загальному випадку розв'язування задачі класифікації даних складається з двох етапів. На першому етапі будують модель, яка описує відому множину класів даних. Таку модель створюють аналізом прикладів (кортежів) відповідної бази даних досліджуваної предметної галузі. Для кожного прикладу визначають клас, якому цей приклад належить, і цей клас позначають певною міткою класу.

На другому етапі побудовану модель використовують для класифікації. Оцінюють точність прогнозування моделлю нових прикладів, для яких відоме значення класифікаційної ознаки. Якщо точність моделі вважають прийнятною, то її можна надалі використовувати для класифікації прикладів, для яких мітка класу невідома.

Аналіз останніх досліджень

До широкого кола задач класифікації широко застосовують дерева рішень і як алгоритми їхнього розв'язування, і як способи відображення знань. Для побудови дерев рішень останніми роками активно застосовують алгоритми CART [5] та ID3 з їх модифікаціями C4.5 та See5 [6, 7]. Всі ці алгоритми будують дерева та дозволяють генерувати правила на основі прикладів.

Таблиця 1

Зразок заповнення таблиці прийняття рішень для визначення кардіологічного захворювання

AGE	GENDER	PIK	KHK_AKMK	KV	SK	UA	AA	BE	OH	REW	R_AK	R_MK	R_AKMK	GH	KHKS
53	M	0	0	0	0	0	0	0	0	0	0	0	0	1	1
65	M	1	0	0	0	0	0	0	0	1	0	0	1	0	0
63	F	1	1	0	0	0	0	0	0	0	0	0	0	1	1
62	M	1	1	0	0	0	0	0	0	0	0	0	0	1	1

Алгоритм ID3 детально описано у праці [6]. У цій же праці наведено його модифікацію C4.5, яка використовує аналіз інформаційного вмісту дерева прийняття рішень для побудови оптимального дерева в сенсі кількості його вершин та висоти. Алгоритм ID3 будує дерево рішень від кореня. Кореню дерева та кожній його внутрішній вершині відповідає певна перевірка з множини перевірок. Для побудови перевірок використовується той факт, що значення кожного умовного атрибута дає змогу розбити множину прикладів на підмножини, у яких усі приклади мають однакове значення цього атрибута.

Якщо рекурсивно застосовувати цей факт, ставлячи у відповідність кожній внутрішній вершині дерева, яке будують, отриману в результаті розбиття множини, то для кожної з отриманих підмножин будуватиметься нове піддерево. Процес побудови піддерев продовжуватиметься доти, доки в результаті поділу не залишаться підмножини, які будуть містити тільки елементи одного і того ж класу з однаковими значеннями атрибута прийняття рішень.

Кожному класу відповідатиме листок дерева рішень. Оскільки послідовність, у якій вершинам дерев приписують атрибути для перевірки, є важливою для побудови дерева рішень, то результат класифікації залежатиме від атрибутів, що приписані кореню і внутрішнім вершинам дерева.

Наведемо абстрактне означення дерева рішень згідно з [8]. Нехай B – деяка непорожня множина та F – деяка множина функцій, які визначені на B і такі, що набувають значення з множини $E_k = \{0, 1, \dots, k-1\}$. Функції з F називатимемо *перевірками*, а пару $U = (B, F)$ – *системою перевірок*. Систему U називають *скінченною*, якщо множина перевірок F скінченна, та *нескінченною* у протилежному випадку. *Задача над U* – це набір вигляду $z = (v, f_1, \dots, f_m)$, де $f_1, \dots, f_m \in F$, $E_k \rightarrow \omega$ та $\omega = \{0, 1, 2, \dots\}$. Задача z полягає у визначенні за довільним елементом $b \in B$ значення $z(b) = v(f_1(b), \dots, f_m(b))$. *Дерево рішень над U* – це кореневе дерево, кожному листку якого відповідає число з ω – результат роботи дерева рішень, кожній внутрішній вершині – перевірка з F , кожній дузі – число з E_k – значення перевірки, за яким перехід відбувається за цією дугою. Дугам, які виходять з однієї вершини, відповідають попарно різні значення.

Таблиця 2

Дискретизація атрибута AGE для двох вікових діапазонів

спосіб дискретизації	Код вікового діапазону	
	0	1
1	0–22	23–88
2	0–32	33–88
3	0–52	53–88
4	0–72	73–88

У *практичних задачах* множини E_k та ω є змістовними: їхні елементи – це, відповідно, значення перевірок та запропоновані рішення, що відповідають змісту проблеми, яку розглядають.

Згідно з [9] введемо інформаційні об'єкти, які розглядатимемо відповідно до введеного означення дерева рішень. Для цього розглянемо множину даних, заданих таблицею, кожний рядок якої містить характеристики певного об'єкта або явища. Кожний стовпчик позначений змінною, яка є атрибутом, значення якого міститься у таблиці. Така таблиця називається *інформаційною системою*. Формально *інформаційна система* – це пара $D = (W, A)$, де W – непорожня скінченна множина об'єктів, а A – множина *умовних* атрибутів, таких, що $a \in A$. Кортеж значень умовних атрибутів кожного об'єкта на множині W називатимемо *прикладом*. Множину V_a називають *множиною значень a* . *Система прийняття рішень* – це інформаційна система вигляду $D = (W, A \cup \{d\})$, де $d \notin A$ – *атрибут прийняття рішень*. Цей атрибут може набувати декількох значень, однак найчастіше використовують двійкові значення $\{0, 1\}$ або $\{Yes, No\}$. Тепер розглядатимемо задачу класифікації як задачу прийняття рішень, а прийняте рішення – як визначення певного атрибута, який відповідає прийнятому рішенню.

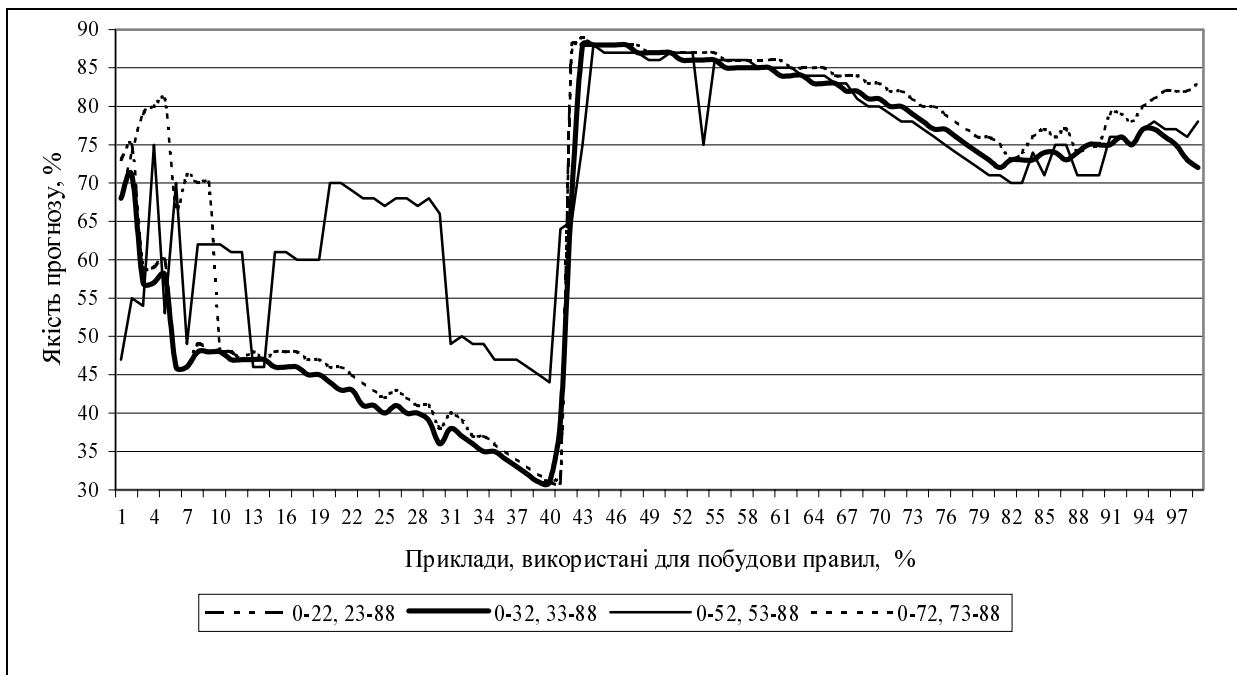


Рис. 1. Прогнозування діагнозу: аналіз якості прогнозу для двох вікових груп.
Експерименти з послідовним вибором навчальних прикладів

У цьому формулюванні задача побудови дерева рішень буде виглядати так: значення атрибута прийняття рішень d визначає певний клас, до якого належить об'єкт з множини W , а множина значень атрибута d виконує розбиття множини W на класи еквівалентності, кожний з яких визначає прийняте рішення. Таблиці прийняття рішень відповідатиме дерево прийняття рішень, яке у цьому випадку є деревом, внутрішнім вершинам якого відповідають тести над атрибутами A . Листки такого дерева відповідають класам еквівалентності, а дугам, які починаються у внутрішніх вершинах дерева, відповідають значення атрибутів, що містяться у цих вершинах.

Наведемо етапи застосування алгоритму ID3 у формулюванні авторів цієї статті [10] при побудові дерева рішень зверху вниз. Ці етапи застосовуються рекурсивно у кореневій та при побудові кожної з внутрішніх вершин дерева прийняття рішень. Процес продовжується доти, поки не буде реалізований етап 1 для всіх вершин дерева. Множина всіх умовних атрибутів позначена A .

Алгоритм ID3

Еман 1. Якщо вершині дерева прийняття рішень відповідає множина прикладів R і всі ці приклади мають однакове значення атрибута прийняття рішень d , то ця вершина є листком дерева і позначається цим значенням d .

Еман 2. Якщо для певної вершини етап 1 не виконується, то розглядається множина умовних атрибутів A . Якщо $A \neq \emptyset$, то вибирається довільний атрибут $a \in A$, яким позначається ця вершина. Атрибут a вилучається з множини A . Для кожного $v \in V_a$ виконуються такі дії:

- утворюють дуги, інцидентні вершині a , кількість яких є $|V_a|$ та позначають кожну з них значенням з V_a ;
- для кожної дуги створюється термінальна вершина;
- множина прикладів R вершини a розбивається на підмножини з однаковим значенням V_a ;
- кожній термінальній вершині відповідає одна з підмножин прикладів, утворених у вершині a , всі елементи якої мають однакові значення V_a .

Еман 3. Якщо на етапі 2 $A = \emptyset$, то щодо вершини, яка повинна стати листком, приймається спеціальне рішення, яке визначається специфікою задачі.

Дерево рішень ставить у відповідність заданому прикладу номер класу. При цьому використовують множину перевірок значень атрибутів цього прикладу у вершинах дерева при обході зверху вниз. (У цій праці термін „обхід” використовуємо для визначення шляху від кореня дерева до певного листка, який і визначає рішення).

Цілі статті

Методика підходу до проектування бази знань використанням дерева рішень із застосуванням алгоритму C4.5 опрацьована на основі реальної інформації, яку надав авторам статті доктор медичних наук, завідувач кафедри терапії Львівського медичного інституту Ю.В. Федоров. Інформація зібрана під час визначення діагнозу певного кардіологічного захворювання. Діагноз визначали аналізом 15 показників, зразок яких наведено у табл. 1.

Таблиця 3

Дискретизація атрибуту AGE для трьох вікових діапазонів

спосіб дискретизації	код вікового діапазону		
	0	1	2
1	0–24	25–52	53–88
2	0–36	37–52	53–88
3	0–40	41–52	53–88
4	0–44	45–52	53–88
5	0–48	49–52	53–88

Для перевірки якості результатів прогнозування таблиця з даними була поділена на дві частини, з яких першу, *навчальну множину*, використовували для побудови дерева рішень (та відповідної бази правил), а другу – для перевірки якості прогнозу. Для цього всі приклади з другої частини таблиці використовувались як *перевірочна множина*, тобто вхідні – для побудованої бази правил, а відомий діагноз порівнювали з тим, що був отриманий застосуванням побудованої бази правил. Інший спосіб перевірки якості прогнозу полягає у використанні всієї множини прикладів заданої таблиці як перевірконої. Саме такий спосіб утворення перевірконої множини були використані при авторах цієї статті.

Зібрані дані використано для вирішення таких завдань:

1. Формування бази знань у вигляді логічних функцій шляхом знаходження прихованих закономірностей у базах даних про пацієнтів та побудова на їх основі системи прийняття рішень для прогнозування діагнозу кардіологічних захворювань.
2. Порівняння впливу способів вибору навчальної множини прикладів для побудови дерева рішень на якість прогнозу. Досліджено два способи вибору таких прикладів. Перший з них здійснює послідовний поділ всієї таблиці на дві частини та утворює множину навчальних прикладів з усієї першої частини таблиці. Другий спосіб полягає у випадковому виборі прикладів, які утворюють навчальну множину прикладів.
3. Оцінка залежності якості прогнозу в групі пацієнтів при різних способах дискретизації неперервного параметра. Таким параметром у наведених даних є вік пацієнта.
4. Зміна прогнозу при збільшенні навчальної множини прикладів.
5. Вироблення рекомендацій щодо організації процесу підготовки даних.

Основний матеріал

Атрибутом прийняття рішень у наведеній табл. 1 є атрибут КНКС, значення “1” якого відповідає наявності хвороби, “0” – її відсутності. Практичною метою досліджень є побудова дерева рішень та на його основі множини правил, з допомогою яких можна автоматизувати процес встановлення діагнозу і цим створити базу знань експертної системи. Таблиця містить результати обстеження 3532 пацієнтів.

**Розподіл пацієнтів за віковими групами
для двох вікових діапазонів**

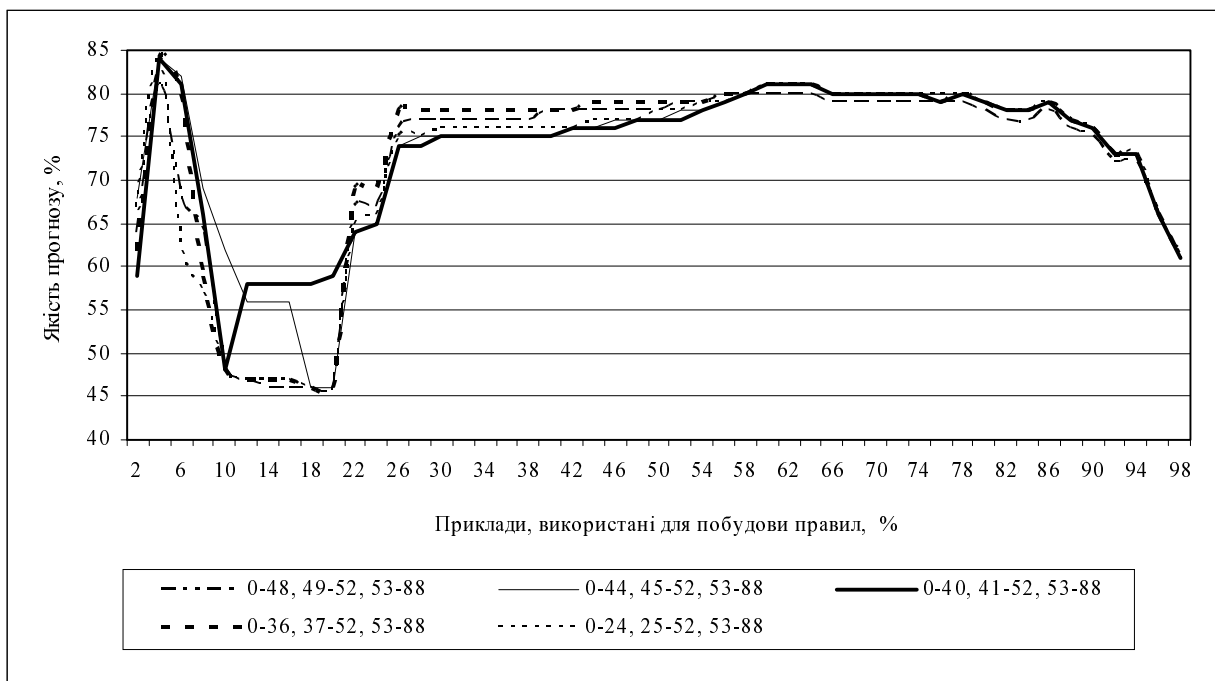
Вік	Разом здорових	% здорових	Разом хворих	% хворих	Разом
0–22	59	100	0	0	59
23–88	2706	77,92	767	22,08	3473
0–32	204	100	0	0	204
33–88	2561	76,95	767	23,05	3328
0–52	1210	93,01	91	6,99	1301
53–88	1555	69,70	676	30,30	2231
0–72	2567	81,05	600	18,95	3167
73–88	198	54,25	167	45,75	365

Для виконання експериментів із застосуванням алгоритму C4.5 побудовано спеціалізоване програмне забезпечення. Дані у таблиці є двійковими, тобто за наявності певного симптому у відповідній клітині ставиться одиниця, а при його відсутності – нуль. Для одноманітності позначень та спрощення застосування процедур аналізу, атрибуту GENDER (стать) надане значення “1” для осіб чоловічої статі та “0” – для жіночої.

Особливістю атрибута AGE (вік) є те, що його значення належать неперервному проміжку із значеннями від 0 до 88. Для застосування процедур аналізу та з огляду на запропоновану медиками специфіку досліджень цей атрибут був дискретизований, тобто весь інтервал його значень розбито на частини, а кожному пацієнту одного вікового діапазону надано одне і те ж значення цього атрибута. Розроблений програмний інструмент дозволяв дискретизувати атрибут AGE, щоби дослідити вплив віку пацієнта на якість прогнозування діагнозу.

За результатами проведених досліджень можна зробити висновок, що випадковий вибір навчальної множини прикладів та розбиття її на вікові групи підвищили якість прогнозу у кожній з них.

Більш глибокі дослідження, пов'язані з проблемою дискретизації та її впливом на якість прогнозування тут не проводились.



*Рис. 2. Прогнозування діагнозу: аналіз якості прогнозу для трьох вікових груп.
Експерименти з послідовним вибором навчальних прикладів*

Під час створення програмного забезпечення врахована проблема суперечливих прикладів. Ця проблема обговорена, зокрема, в [9] та [10]. Такі приклади містять однакові значення всіх умовних атрибутів, але відрізняються значенням атрибута прийняття рішень. Появу у таблиці таких атрибутів можна пояснити двома причинами: або такий атрибут записаний з помилкою, або в процесі прийняття рішення експерт користувався ще додатковими міркуваннями, які не відображені у таблиці. Для уникнення суперечностей, що можуть виникнути у зв'язку з наявністю таких атрибутів при побудові дерева рішень, використано той приклад з множини суперечливих прикладів, кількість представників якого була більшою. Решту прикладів, що мали менше представників, – відкинута.

На графіках, що зображені на рис. 1–6, показані залежності якості прогнозу від кількості прикладів, використаних для побудови дерев рішень та відповідних баз правил. Тут якість прогнозу визначалась як процент прикладів, для яких діагноз, обчислений з допомогою бази правил, збігався із значенням атрибута прийняття рішень. Ефективність застосування побудованих дерев рішень аналізувалася порівнянням двох груп експериментів, які відрізнялись способом утворення навчальної множини прикладів. У першій групі експериментів навчальна множина формувалась з прикладів, що вибирались послідовно з верхньої частини таблиці.

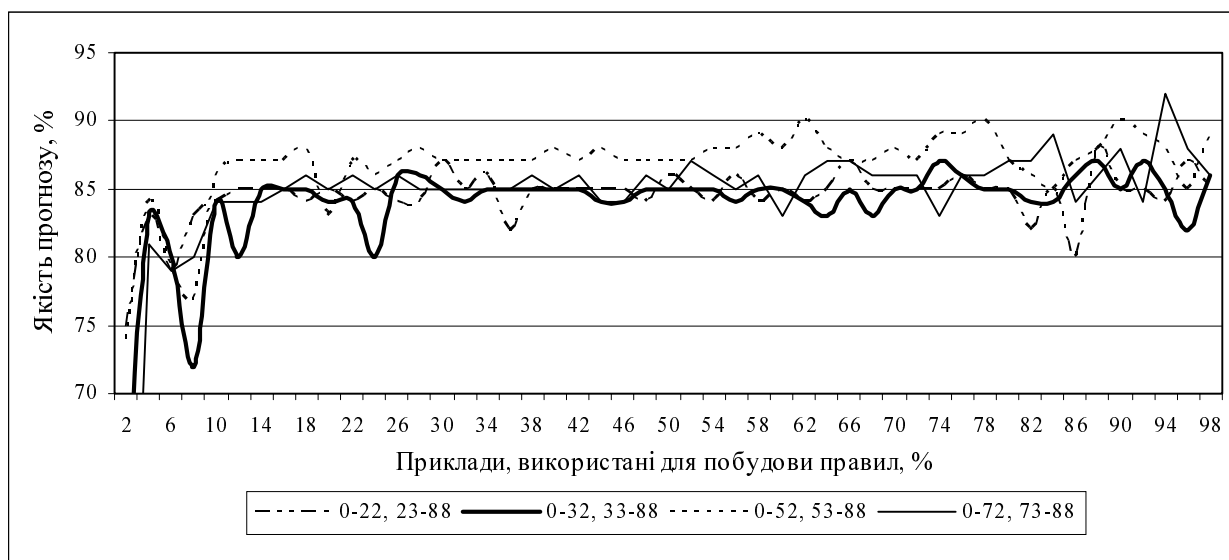


Рис. 3. Прогнозування діагнозу: аналіз якості прогнозу для двох вікових груп. Експерименти з випадковим вибором навчальних прикладів

Залежності якості прогнозу від відсотка прикладів, використаних для побудови правил, показані на рис. 1 та 3, відповідають чотирьом способам дискретизації вікового діапазону (атрибут AGE) на дві групи, молодша з яких позначена символом „0”, а старша – символом „1”. Ці способи дискретизації розбивають пацієнтів на групи, спосіб утворення яких ілюструє табл. 2. Розподіл пацієнтів на ці вікові групи наведено у табл. 3. Графіки на рис. 2 та 5 відповідають п'яти способам дискретизації вікового діапазону при розбитті цього діапазону на три вікові групи, які позначено як 0, 1 та 2, відповідно, а спосіб утворення цих груп ілюструє табл. 4. Розподіл пацієнтів на ці вікові групи показано у табл. 5. Графіки, зображені на рис.1 та 2, відповідають послідовному вибору навчальних прикладів, а на рис.3,4, 5 та 6 – випадковому.

При обох способах дискретизації на рис.1 та на рис. 2 можна виділити три зони. У першій зоні, яка охоплює значення від 0 до приблизно 43% прикладів на рис.1 та приблизно до 25% на рис.2, при збільшенні кількості навчальних прикладів помітно зменшується якість прогнозу; у другій зоні в проміжку від 43% до приблизно 70% на рис. 1 та від 25% до приблизно 80% на рис. 2 відносно стабілізується якість прогнозу; у третій зоні після 70% прикладів на рис. 1 та після 80% на рис. 2 знову спостерігається падіння якості прогнозу.

Поясненням такої закономірності якості прогнозу є наявність особливостей організації збирання інформації та встановлення діагнозу, яка явно не відображена у таблиці. Лише в результаті спілкування з лікарями, причетними до формування даної таблиці прийняття рішень, з'ясовано, що процес її створення тривав впродовж багатьох років та пройшов три стадії.

Таблиця 5

Розподіл пацієнтів на вікові групи для трьох вікових діапазонів

Вік	Разом здорових	% здорових	Разом хворих	% хворих	Разом
0–24	80	100	0	0,00	80
25–52	1130	92,55	91	7,45	1221
53–88	1555	69,70	676	30,30	2231
0–36	289	99,66	1	0,34	290
37–52	921	91,10	90	8,90	1011
53–88	1555	69,70	676	30,30	2231
0–40	460	98,92	5	1,08	465
41–52	750	89,71	86	10,29	836
53–88	1555	69,70	676	30,30	2231
0–44	634	97,54	16	2,46	650
45–52	576	88,48	75	11,52	651
53–88	1555	69,70	676	30,30	2231
0–48	877	96,06	36	3,94	913
49–52	333	85,82	55	14,18	388
53–88	1555	69,70	676	30,30	2231

На першій стадії інформація накопичувалась у відділі функціональної діагностики при загальному обстеженні кардіологічних хворих. Таке обстеження мало скринінговий характер, тобто не приділялось спеціальної уваги окремим показникам здоров'я пацієнта та можливим причинам виникнення досліджуваної кардіологічної хвороби (кальціонуючої хвороби клапанів серця).



Рис. 4. Прогнозування діагнозу: усереднений прогноз за двома віковими групами. Експерименти з випадковим вибором навчальних прикладів

На другій стадії обстеження набуло моніторингового характеру, тобто було виділено групу з 15 показників, на які зверталась особлива увага при обстеженні хворих. На третій стадії, коли було вже обстежено значну кількість пацієнтів і встановлено загальні закономірності виникнення хвороби, знову повернулись до скринінгової методики досліджень, тобто вказаним показникам знову приділялась увага на рівні інших ознак. Саме ці три стадії проявились на графіках, що наведені на рис.2 та 3.

Для того, щоб зменшити вплив вказаної залежності, запропоновано утворювати навчальну множину випадковим вибором прикладів з рівномірним розподілом. Множини навчальних прикладів утворювались для дискретизації атрибута AGE, відображеного у табл. 2 та 3. На рис. 3 та 5 показано залежності якості прогнозу від розміру навчальної множини прикладів при такому способі її утворення.

На обох графіках помітні дві зони: у першій з них для розміру навчальної множини від 0 до приблизно 15% всіх прикладів зростає якість прогнозу, у другій зоні – якість прогнозу утримується у достатньо вузькому діапазоні. Як вже було вказано раніше у цій статті, дискретизація вікового параметра більшої кількості вікових діапазонів та утворення трьох вікових груп замість двох підвищили якість прогнозу в кожній з цих груп. На рис. 4 та 6 показано графіки середніх значень прогнозів, що відповідають графікам з рис. 3 та 5.

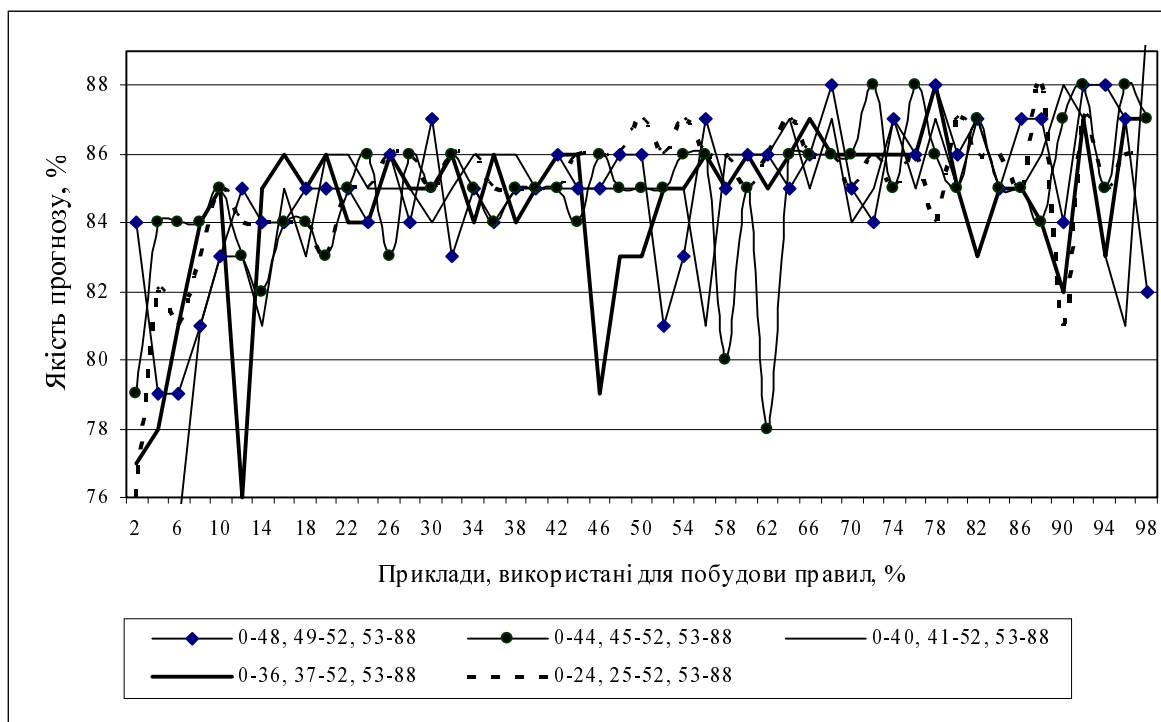


Рис. 5. Прогнозування діагнозу: аналіз якості прогнозу для трьох вікових груп. Експерименти з випадковим вибором навчальних прикладів

Висновки

Результатом проведених досліджень є побудована база знань у формі логічних функцій на основі знайдених прихованих закономірностей у базі даних про пацієнтів. База знань використана для побудови системи прийняття рішень для прогнозування діагнозу кардіологічних захворювань. При використанні створеного програмного забезпечення вдалось встановити залежність якості прогнозу від способу побудови навчальних прикладів. Випадковий вибір прикладів дозволив забезпечити зростання якості прогнозу при збільшенні кількості прикладів. Послідовний спосіб вибору прикладів за зростанням дати запису у базу даних дозволив помітити закономірності організації процесу підготовки інформації та використати ці закономірності для вироблення рекомендацій щодо цього процесу.

Оцінено вплив на якість прогнозу різних способів дискретизації неперервного параметра, яким є вік пацієнта. Помічено, що спосіб дискретизації впливає на якість рішень, але оптимізація процесу дискретизації залишилась за межами проведених досліджень та вимагає додаткового вивчення.

Важливою проблемою, яка вимагає окремого вивчення, є дослідження впливу способів усунення суперечностей та невизначеностей, які існують у реальних даних, на якість прогнозу.

Усуненням таких невизначеностей можна підвищити якість прогнозу, знайти нові закономірності у даних або зменшити розміри таблиці прийняття рішень вилученням однакових рядків та таких стовпців з таблиці, які не впливають на прийняте рішення.



Рис. 6. Прогнозування діагнозу: усереднений прогноз за трьома віковими групами. Друга група експериментів з випадковим вибором навчальних прикладів

Автори статті висловлюють подяку Р.Т. Якимечку за проведений обчислювальний експеримент.

1. Джексон П. Введение в экспертные системы. – М.: Издательский дом «Вильямс», 2001.
2. Mitchell T. Machine Learning. The McGraw-Hill Companies, Inc. 1997.
3. Czichosz P. Systemy uczasa sie. Warszawa: Wydawnictwa Naukowo-Techniczne, 2000.
4. Pyle D. Data Preparation for Data Mining. Morgan Kaufmann Publisher, 1999.
5. Duda R.O., Hart P.E., Stork D.G. Pattern Classification. John Wiley & Sons, Inc. 2000.
6. Quinlan R. C.4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.
7. Dunham M. Data Mining Introductory and Advanced Topics. Pearson Education Inc., 2003.
8. Мошков М.Ю. Деревья решений. Теория и приложения. – Нижний Новгород: Изд-во Нижегород. ун-та, 1994.
9. Komorowski J., Pawlak P., Polkowski L. and Skowron A. Rough Sets: A Tutorial // Eds. S.K.Pal and A. Skowron, Rough Fuzzy Hybridization: A New Trend in Decision-Making, Springer-Verlag, Singapore, 1998. – P.3–98.
10. Никольський Ю.В., Щербина Ю.М., Якимечко Р.Я. Деревя прийняття рішень та їхнє застосування для прогнозування діагнозу у медицині // Вісник Львів.ун-ту. Серія прикл.мат. та інформатика. – 2003. – Вип. 6. – С.191–211.