основі об'єктно-орієнтованої моделі представлення знань // Вісн. Нац. ун-ту "Львівська політехніка". – 2001. – № 433. – С. 94–102. 4. Логический подход к искусственному интеллекту: от классической логики к логическому программированию / А. Тейз, П. Грибомон, Ж. Луи и др. – М., 1990. 5. Литвин В.В., Кравець Р.Б. Інтелектуальні інформаційні системи з кількома інтерпретаторами // Вісн. Нац. ун-ту "Львівська політехніка". – 2001. – № 438. – С. 104–108. 6. Калянов Г.Н. CASE: Структурный системный анализ. – М., 1996.

УДК 681.3

Ю.В. Нікольський
Національний університет "Львівська політехніка",
кафедра "Інформаційні системи та мережі"

ЗАСТОСУВАННЯ МЕТОДІВ КЛАСТЕРНОГО АНАЛІЗУ ПРИ ПОБУДОВІ КЛАСИФІКУЮЧИХ ПРАВИЛ В ЗАДАЧІ ПРИЙНЯТТЯ РІШЕНЬ[1]

The application the cluster analysis method in constructing the classificators for medical diagnosis is proposed. Classificator is marking by means of associated rules. Cluster analyse as a special heuristics allows decreasing the number of features in the left part of the rule. The method is used for creation the rules from decision table that was obtained for the diagnosis of some cardiac decease.

Застосовано метод кластерного аналізу при побудові класифікатора для встановлення медичного діагнозу. Класифікатор будується за допомогою асоціативних правил. Кластерний аналіз, який використано як спеціальну евристику, дозволяє скоротити кількість ознак у лівій частині правила. Метод використано для створення правил з таблиці прийняття рішень, побудованої для діагностування певного кардіологічного захворювання

ВСТУП

Характерною особливістю суспільних процесів та наукових досліджень останнього десятиріччя є широке застосування інформаційних технологій. Це призвело до накопичення колосальних обсягів інформації, обсяги якої продовжують зростати в експоненційному порядку. Проблема використання накопичених даних лежить у площині пошуку способів аналізу цієї інформації, отримання з цих даних корисних залежностей, які можуть бути застосовані для прийняття рішень.

Розвиток штучного інтелекту в останні роки зорієнтований на побудову систем, основаних на побудові баз знань із застосуванням індуктивних технологій. Загальна тенденція цих досліджень полягає у знаходженні прихованих закономірностей у даних та

[1] Дослідження, результати яких подані у цій статті, частково підтримані ґрантом від бюро з питань освіти та культури (ECA) Держдепартаменту США. ECA не несе відповідальності за погляди, що тут висловлені.

формування на їх основі моделей систем прийняття рішень у вигляді систем вирішальних правил. Такі правила розв'язують задачу прийняття рішень як задачу класифікації шляхом віднесення об'єкта, щодо якого приймається рішення, до певної групи об'єктів наданням йому певної класифікуючої ознаки. Існує погляд, що вказані класи задач належать до більш широкого класу, вважаючи опис кожного об'єкта його образом, а задачі класифікації задачами розпізнавання цих образів [9, 10].

За своєю суттю моделі систем прийняття рішень як системи класифікуючих правил є логічними функціями певного вигляду, а задача класифікації полягає в обчисленні цих функцій. Вказані функції повинні відображати реально існуючі закономірності та давати можливість приймати рішення, що мають високу точність. Це означає, що до побудованих моделей ставиться вимога отримання результату, що мало відрізняється від рішень, прийнятих експертами. Важливою проблемою, яка виникає при побудові таких моделей, є повнота даних щодо реальної ситуації. При цьому треба мати на увазі, що при збиранні реальної інформації значна частина даних може бути неточною, суперечливою або взагалі відсутньою.

Для побудови вказаних систем застосовують різні адаптаційні алгоритми, які називають ще методами *машинного навчання*. Виділяють два основні підходи, що отримали назви *"навчання з вчителем"(supervised lesrning)* та *"навчання без вчителя"(unsupervised learning)*. При першому підході процес класифікації орієнтується на відомі рішення, що здійснені експертом, а в другому – отримання груп об'єктів здійснюється шляхом пошуку спільних їх рис. Такі об'єкти мають назву *кластерів*, а утворюють їх методами *кластеризації*.

АНАЛІЗ ОСТАННІХ ДОСЛІДЖЕНЬ

Системи підтримки прийняття рішень будуються знаходженням правил, отриманих з таблиць даних. Такі системи класифікують дані, а способи їх побудови відрізняються алгоритмами отримання правил та формами, у яких ці правила записуються. Всі можливі підходи вирішення задачі побудови систем правил прийняття рішень об'єднують питання, на які ці системи дають відповідь. Найважливіше з них полягає у можливості класифікації цими правилами нових об'єктів.

Проблематика побудови систем класифікуючих правил у медицині розглядається багатьма авторами, зокрема [8]. Зазначається, що основними задачами аналізу медичних даних, які мають відношення як до постановки діагнозу, так і призначення лікування, є такі:

- визначення найважливіших ознак, тобто спеціальних характеристик пацієнта для їх класифікації з точки зору діагностики;
- знаходження взаємозв'язку між значеннями найважливіших ознак та класифікацією пацієнтів.

Для побудови класифікуючих правил існує велика кількість спеціальних алгоритмів [6]. З обчислювальної точки зору ці алгоритми є *NP* -складними, і їх застосування вимагає використання спеціальних евристик для зменшення загального обсягу обчислень. При цьому загальна кількість отриманих правил може бути значною, що вимагатиме додаткових зусиль для вибору кращих з них. Серед відомих підходів до формування таких евристик є метод булевих міркувань *(Boolean reasoning)* [11].

Вирішення задачі прийняття рішень (класифікації) як проблема встановлення діагнозу з точки зору застосування отриманих правил повинна спиратись на досвід діагностування

в умовах конкретного медичного закладу. Слід зауважити, що у медичній практиці кількість можливих симптомів для встановлення діагнозу є невеликою, тому проблема знаходження правил полягає у визначенні таких основних ознак, за якими найточніше вирішується задача постановки діагнозу.

Пропонується будувати евристику, основану на застосуванні методу кластерного аналізу для скорочення кількості ознак, які будуть використані при подальшій побудові правил в задачі діагностування.

Для подальших досліджень розглядатимемо вихідну таблицю даних про об'єкти як таблицю прийняття рішень [7] у вигляді пари $B=(U, A\cup\{d\})$. Тут U – множина всіх об'єктів, а $A=\{a_1, a_2, ..., a_m\}$ – множина атрибутів, які встановлюють відповідність вигляду $a_i : U \to V_a$, де V_a – множина значень атрибуту a_i, d – *атрибут прийняття рішень*. Атрибути множини A називаються *умовними* або умовами, а d – рішенням. Припускається, що множина d має скінченний розмір $rank(d)$. Множина об'єктів $C_i=\{o\in U : d(o)=d_i\}$ є i-м класом рішень, d_i – значення i-го рішення, отриманим з множини рішень $V_d=\{d_1, d_2, ..., d_{rank(d)}\}$. Правило прийняття рішення – це формула вигляду $(a_{i_1}=v_1)\wedge(a_{i_2}=v_2)\wedge...\wedge(a_{i_k}=v_k)\Rightarrow d=v_d$, де $1\le i_1<...<i_k\le m$, $v_i\in V_{a_i}$. Атомарна формула $(a_{i_1}=v_1)$ називається умовою. Кажемо, що правило r *застосовне* до об'єкта, або, навпаки, об'єкт задовольняє правило, якщо значення його атрибутів задовольняє припущення правила $p=(a_{i_1}=v_1)\wedge(a_{i_2}=v_2)\wedge...\wedge(a_{i_k}=v_k)$. З правилом пов'язуються певні числові характеристики. *Підтримка* (*support*), що позначається $Supp_B(r)$, дорівнює кількості об'єктів з B, для яких правило r коректно застосовується, тобто задовольняється припущення правила та рішення, що отримується за правилом, відповідає значенню, що міститься у таблиці. *Довіра (confidence)* $Conf_B(r)$ визначається як відносна кількість коректно застосованих правил, для яких задовольняються припущення, та обчислюється за формулою $Conf_B(r)=\dfrac{Supp(p\Rightarrow d=v_d)}{Supp(p)}$.

Для аналізу медичних даних, який орієнтований на пошук закономірностей утворення груп пацієнтів, використано технологію кластерного аналізу. Застосування методів кластерного аналізу можна розглядати з точки зору попередього поділу груп даних з метою виділення визначальних ознак, за якими можна розрізнити групи пацієнтів. Ці ознаки виділяються з множини ознак, які є певними симптомами або результатами аналізів і використовуються на етапи попереднього обстеження хворого, що передує встановленню діагноза і призначенню лікування.

Кластеризація полягає у побудові груп об'єктів, які є у певному сенсі подібними між собою. На відміну від класифікації, ці групи не мають попередньо визначених ознак (прийнятих рішень), до яких може бути віднесений кожний з цих об'єктів. Групування відбувається знаходженням подібності між об'єктами на основі властивостей, що містять ці об'єкти. Отримані групи називаються кластерами. Існують різні означення кластера, які мають основні спільні моменти:

Кластер – множина подібних об'єктів, а елементи різних кластерів відмінні між собою.

Відстань між елементами у кластері є меншою від відстані між елементом у кластері та довільним елементом за його межами.

Ідея кластеризації об’єктів, очевидно, поширюється на сегментацію бази даних, при якій групуються кортежі або записи бази даних про вказані об’єкти. Це дозволяє сегментувати базу даних на групи, що дає досліднику можливість побачити загальні властивості об’єктів. Тому в цьому контексті можна казати про сегментацію та кластеризацію як про синоніми.

При застосуванні процесу кластеризації враховуються такі його особливості:

1. Наперед невідома кількість кластерів, які будуть утворюватись.
2. Немає жодних апріорних знань, що стосуються утворюваних кластерів.

Для загальної постановки задачі кластеризації припускаємо, що кількість кластерів є вхідним параметром l. Поточне значення кожного кластера K_j, $1 \le j \le l$ визначене як результат застосування деякої функції. Без втрати загальності результатом розв’язування задачі кластеризації є створення множини кластерів $K = \{ K_1, K_2, ..., K_l \}$.

Для заданої множини кортежів бази даних $D = \{ t_1, t_2, ..., t_n \}$ та цілого числа l задача кластеризації полягає у знаходженні відповідності $f : D \to \{ 1, ..., l \}$, де кожне d_i відповідає одному кластеру K_j, $1 \le j \le l$. Кластер K_j містить тільки ті кортежі, що йому відповідають, тобто $K_j = \{ d_i \mid f(d_i) = K_j,\ 1 \le i \le m \wedge d_i \in D \}$.

Існує кілька підходів до вирішення задачі кластеризації. У цій статті використано метод ієрархічної кластеризації. Результат застосування методу ієрархічної кластеризації зображають повним бінарним деревом – *дендрограмою*. Методи ієрархічної кластеризації поділяються на *подільні (divisive)* та методи *концентрації (agglomerative)* [2], відповідно до стратегії формування кластерів – ”згори-донизу” або “знизу-догори”.

У стратегії “знизу-догори” кожний об’єкт на початку кластеризації відноситься до окремого кластера, а у процесі кластеризації відбувається об’єднання кластерів у більші кластери. Цей процес продовжується доти, поки всі об’єкти не опиняться в одному кластері або не буде виконаний певний критерій зупинки. Більшість відомих алгоритмів реалізують саме цю стратегію. Вони відрізняються між собою у способі визначення кластерів, які належить об’єднувати на кожному кроці алгоритму за ознакою міжкластерної подібності або близькості кластерів.

Алгоритми ієрархічної кластерної подільності, що реалізують стратегію “згори-донизу”, діють протилежно до ієрархічної концентрації. Вони починаються від одного кластера, що містить всі об’єкти. Цей метод розбиває кластери на все менші частини, поки кожний кластер не складатиметься з одного об’єкта або не будуть виконані інші умови зупинки, такі як отримання певної кількості кластерів або досягнення певної величини відстані між двома найближчими кластерами. Прикладами методів вказаних класів є метод концентрації AGNES (Agglomerative NESting) та метод подільності DIANA (Divisive ANAlysis) [3]. На рис. 1 [2] показано кластеризацію множини об’єктів $\{a, b, c, d, e\}$, на якому дендрограма зображає процеси об’єднання (переглядом зліва направо) та поділу (переглядом справа наліво) кластерів.

Пошук закономірностей у даних виконуємо за методом ієрархічної концентрації. Процес побудови дендрограми починається з розгляду наперед визначеної множини, що складається з n об’єктів. На першому кроці алгоритму кожен об’єкт із заданої множини вважається одним кластером. Два найближчі об’єкти об’єднуються в один кластер, і їх загальне число тепер дорівнюватиме $n-1$. Близькість об’єктів обчислюється за формулою,

яка є фукцією відстані. Серед кластерів, що залишились, знову відшукуються найближчі, які також об'єднуються. Така процедура продовжується доти, поки всі об'єкти не потраплять в один кластер.

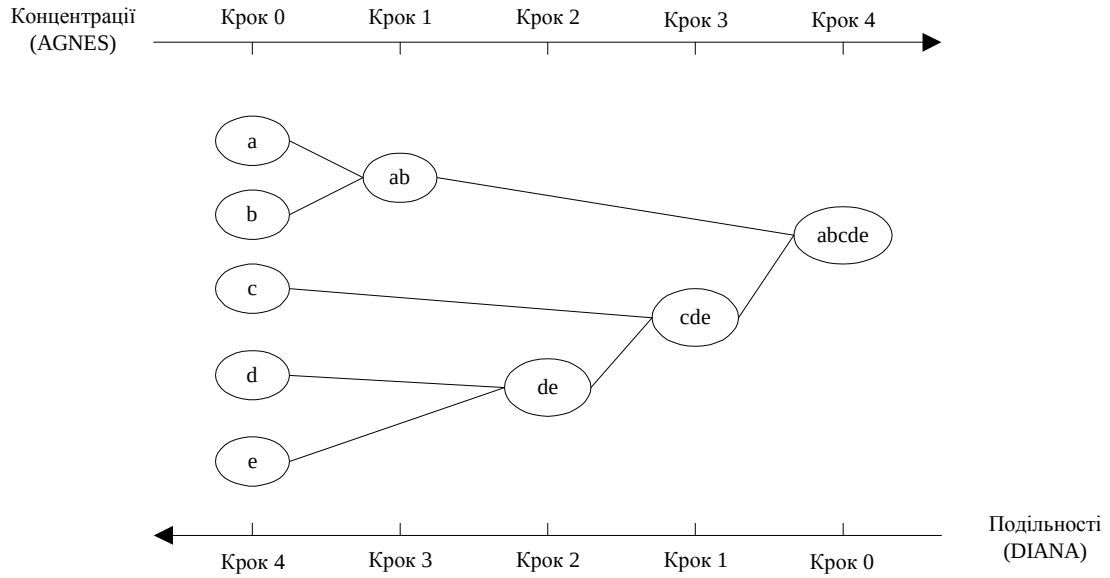

Рис. 1. Ілюстрація підходів до реалізації процесу кластеризації

ФОРМУВАННЯ ЦІЛЕЙ

Об'єктом досліджень є таблиця прийняття рішень, кожний рядок якої містить перелік властивостей певного явища або об'єкта. Аналіз полягає у сегментації таблиці з метою пошуку груп типових об'єктів та визначенні спільних закономірностей об'єднання об'єктів у групи. У пропонованому дослідженні цей аналіз виконувався у три етапи:

- *на першому етапі* знайдено відстані між всіма парами об'єктів та побудовано матрицю відстаней. Відстань між двома об'єктами, що є векторами $t_i = (t_{i1}, t_{i2}, ..., t_{ik})^T$ та $t_j = (t_{j1}, t_{j2}, ..., t_{jk})^T$, визначалась за правилом $d(t_i, t_j) = \sum_{h=1}^{k} | t_{ih} - t_{jh} |$. Тут t_{ij} – j-ва ознака i-го кластера, $i = 1,2,...,n$, $j = 1,2,...,k$. Така відстань називається *манхеттенською*.
- *на другому етапі* застосовано алгоритм агломеративної ієрархічної кластеризації із скороченням матриці відстаней за методом медіани та об'єднанням в один кластер таких, відстань між якими є найменшою. За методом ієрархічної кластеризації побудована дендрограма.
- *на третьому етапі* проводився аналіз дендрограми з метою знаходження закономірностей об'єднання кластерів та формулювання відповідних правил, які відображають характер розподілу значень атрибутів в процесі кластеризації. Аналіз проводиться шляхом обходу дерева, що зображає дендрограму, від кореневої вершини в напрямку його внутрішніх вершин. У кожній вершині аналізуються її кластери у вершинах-синах, з точки зору перерозподілу між ними значень ознак.

Детальніше розглянемо виконання вказаних етапів аналізу. Для побудови дендрограм використано матрицю відстаней. Відомі множина об'єктів $S = \{ S_1, S_2, \dots S_n \}$ та множина ознак $Z = \{ Z_1, Z_2, \dots, Z_q \}$. Кожна з ознак набуває значень зі свого домена. Кожний об'єкт задається кортежем $D_i = \{ t_{1i}, t_{2i}, ..., t_{qi} \}$, $i = 1,...,n$ значень своїх ознак та утворює відповідний рядок таблиці, а кожний об'єкт утворює кластер $H_i, i = 1,2...,n$.

Алгоритм ієрархічної кластеризації:

Крок 1. Для всіх кортежів вигляду $D_i = \{ t_{1i}, t_{2i}, ..., t_{qi} \}$, $i = 1,...,n$ будується квадратна матриця з елементами $d(H_i, H_j)$, $i, j = 1,...,n$ відстаней між усіма парами об'єктів. Ця матриця має розміри $p \times p$, де $p = \binom{n}{2} = \frac{n(n-1)}{2}$.

Крок 2. Серед елементів матриці відстаней вибираються елементи з мінімальним значенням; нехай це d_{ul}. Якщо у матриці є більше відстаней, ніж один елемент з таким значенням, то вибирають довільний з них, можливо, випадково.

Крок 3. Рядок з номером u та стовпець з номером l вилучають з матриці відстаней, а кластери H_u та H_l об'єднують у кластер H_k.

Крок 4. Перераховують всі відстані $d(H_j, H_k)$, $H_j \neq H_u$, $H_j \neq H_l$ за *методом медіани*. Для цього відстані $d(H_j, H_k)$ між кластером H_k та довільним кластером H_j перераховують за формулою $d(H_j, H_k) = 0{,}5\left(d(H_j, H_u) + d(H_j, H_l) \right)$. Отримана матриця відстаней матиме розміри $(p-1) \times (p-1)$.

Крок 5. Якщо матриця відстаней зменшиться до розмірів 2×2, то процес ієрархічної кластеризації закінчується, оскільки отримано два кластери, які можливо об'єднати в один кластер, що містить всі об'єкти. Інакше – повернутись до кроку 2.

Результатом виконання алгоритму є кластери та відстані між ними у послідовності їх утворення, що зображається дендрограмою.

ОСНОВНИЙ МАТЕРІАЛ

У цій частині обговорюється запропонований підхід до аналізу бази даних, які зібрані при встановленні діагнозу реального кардіологічного захворювання. Ці дані досліджувались іншими методами у працях [4, 5]. Для аналізу використано дані про 3532 пацієнтів. Ствопці таблиці ознак є такими: GENDER, R_AK, PIK, KV, SK, BE, OH, AA, REW, UA, R_MK, GH, R_AKMK, AGE25, AGE20_40, AGE35_55, AGE50, KHKS. Кожна ознака набуває значення 0 або 1. Ознака GENDER має значення 1, якщо пацієнт – жінка, та 0, якщо – чоловік. Стовпці таблиці мають наступний зміст та значення:

- KHKS – діагноз захворювання із значенням 1 для хворих та 0 – для здорових;
- AGE25 – вік пацієнта із значенням 1 для пацієнтів, молодших за 25 років та 0 – для старших за 25 років;
- AGE20_40 – вік пацієнта із значенням 1 для пацієнтів у віці від 20 до 40 років, та 0 – для пацієнтів іншого віку;
- AGE35_55 – вік пацієнта із значенням 1 для пацієнтів у віці від 35 до 50 років, та 0 – для пацієнтів іншого віку;
- AGE50 – вік пацієнта із значенням 1 для пацієнтів, старших від 50 років та 0 – для пацієнтів іншого віку.

Значення решти властивостей є 0 або 1, що відповідає відсутності або наявності результатів вказаних аналізів або симптомів.

Надалі замість терміну “кластер” будемо вживати вираз “група пацієнтів”. Результати другого етапу аналізу зображено на рис. 2 фрагментом дендрограми, яка має вигляд повного бінарного дерева з позначеними вершинами. Тут показано останні 14 кроків побудови дендрограми, де N – кількість елементів, які аналізуються у даній вершині, M – кількість хворих, які потрапили до цієї групи, D – відстань між кластерами, які об’єднано у цій вершині та k – номер вершини.

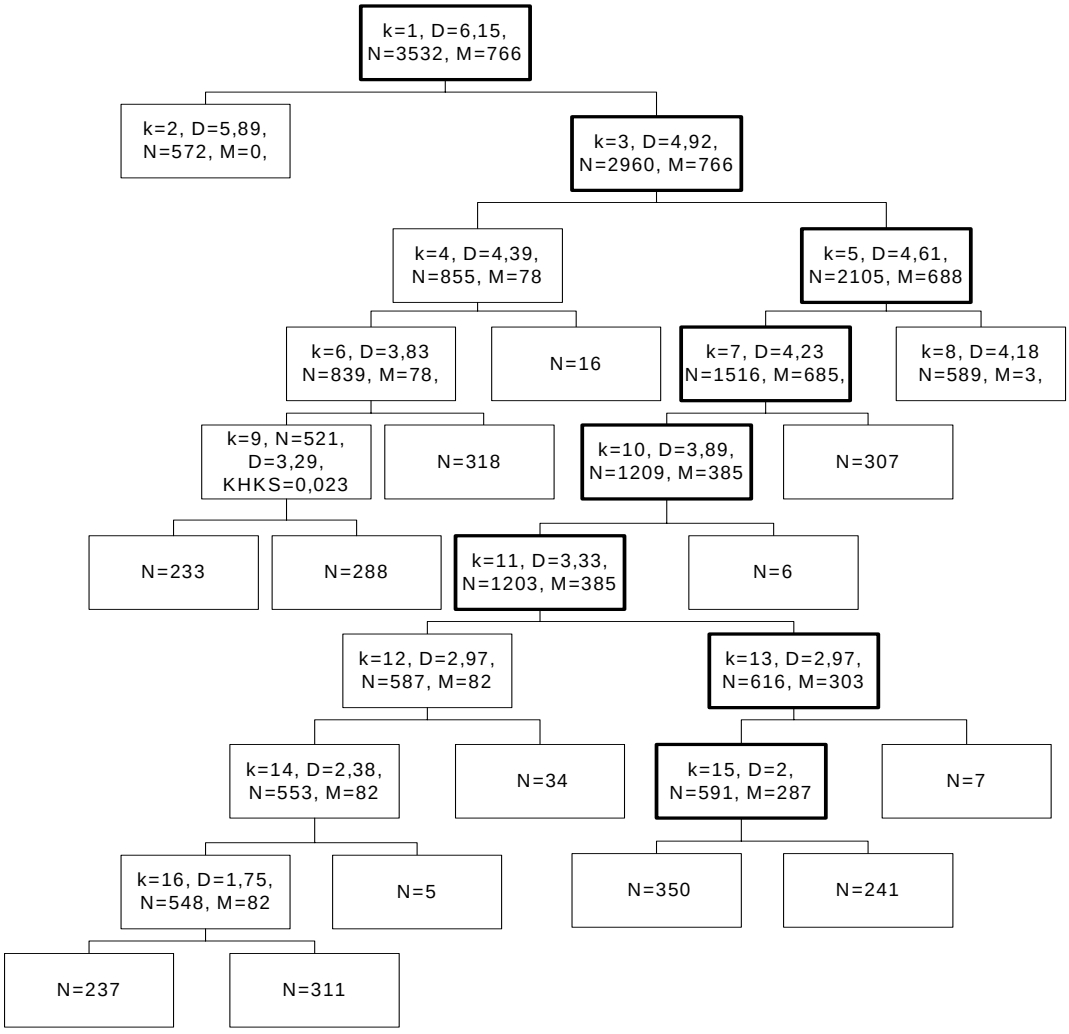

Рис. 2. Фрагмент дендрограми, утвореної застосуванням кластерного аналізу до таблиці даних про встановлення діагнозу кардіозахворювання

Легко зауважити, що відстань між кластерами збільшується при просуванні згори донизу по дереву. Аналіз процесу поділу на групи припинено в момент отримання групи, відсоток хворих в якій є меншою ніж 5 % загальної кількості всіх хворих, тобто при $M \le 38$.

Результати формування груп формулюються на третьому етапі аналізу правилами вигляду: *(R%) IF* $A = S$ *TO GROUP* L *(E%)*, де A – ім’я ознаки, S – значення ознаки A, L – номер групи ($L = 1$ для групи з $N1$ пацієнтами, а $L = 2$ – з $N2$ пацієнтами), R – відсоток пацієнтів з $A = S$ у групі з номером L, E – відсоток всіх пацієнтів з $A = S$ у групі з

номером L. Всі пацієнти перед останнім об'єднанням містились у групах з N1=2960 та N2=572 осіб.

Табл. 1 містить інформацію, що відповідає вершинам графа з номерами 1, 2 та 3 ($k = 1,2,3$). У рядках цієї таблиці містяться значення ознак у групах з кількостями пацієнтів відповідно 3532, 572 та 2960. З таблиці також видно, що у першій групі зросла відносна кількість елементів, що містить значення PIK, KV, AGE50 та GH. Значенню ознак відповідає відсоток пацієнтів, які мають ці ознаки. Для наочності на рис. 4–6 показано діаграми розподілу ознак з таблиці, а на рис. 3 – порівняння відносних значеннь цих ознак. Крім того, в останніх рядках таблиці міститься інформація про підгрупи, на які поділяється кожна з груп. У табл. 2 додатково показано розподіл пацієнтів за кожною з ознак. Так, у рядку, позначеному символом "0", міститься розподіл пацієнтів, у яких значення відповідної ознаки дорівнює нулю, між підгрупами після поділу груп. Аналогічні дані містяться у рядках, позначених символом "1". Дані у табл. 1 та 2 дозволяють зробити такі висновки: серед усіх пацієнтів є 21,7 % хворих (KHKS=0,21716, табл. 1).

Таблиця 1

Значення ознак пацієнтів, отриманих при останньому об'єднанні підгруп

K	1	2	3	K	1	2	3
N	3532	572	2960	R_MK	0.02350	0.14500	0
GENDER	0.62316	0.57200	0.63311	GH	0.45527	0.21000	0.50270
R_AK	0.00566	0.03500	0	R_AKMK	0.01812	0.11200	0
PIK	0.42016	0.12600	0.47703	AGE25	0.02548	0.15700	0
KV	0.15317	0.00699	0.18142	AGE20_40	0.12684	0.74300	0.00777
SK	0.01869	0.01050	0.02027	AGE35_55	0.37627	0.46000	0.36014
BE	0.00481	0.00175	0.00541	AGE50	0.71263	0.17700	0.81622
OH	0.00198	0	0.00236	KHKS	0.21716	0.00175	0.25878
AA	0.00255	0	0.00304	D	6.15466	5.89000	4.92374
REW	0.04700	0.29000	0	N1	2960	159	2105
UA	0.02916	0.00699	0.03345	N2	572	413	855

Зокрема, правило 2 означає, що 81,6 % всіх пацієнтів першої групи старші за 50 років, причому 96 % всіх пацієнтів цього віку потрапили саме до першої групи. Якщо розглядати розподіл хворих (за ознакою KHKS), то вони становлять 25,9 % всіх, хто є у першій групі, разом з тим у цю групу потрапило 99,9 % всіх хворих. Тепер можна зробити висновки щодо розподілу інших ознак. Так, у першій групі є 47,7 % пацієнтів, які мають ознаку PIK=1 та становлять більше ніж 95 % всіх пацієнтів з цим значенням атрибуту PIK. Крім того, у цій групі 50,5 % пацієнтів мають GH=1 та становлять 92,5 % всіх пацієнтів з цим значенням ознаки GH. Щодо значення інших ознак, можна зробити відповідні висновки.

У першій групі є 25,8 % хворих, у другій – 0,17 %. Групи пацієнтів описані правилами розподілу ознак, наведеними у табл. 3. Зазначимо, що з точки зору встановлення причинно-наслідкових зв'язків для визначення факторів, що викликають захворювання, доцільно розглядати пацієнтів, що потрапили до першої групи. Внаслідок встановлення системи правил для задачі, що розв'язується, отримуємо такі результати. Атрибутом прийняття рішення у випадку задачі, що розглядається, є ознака KHKS $(d = KHKS)$. Решта ознак є умовними атрибутами: $V_d = \{0,1\}$, $rank(d) = 2$. Множина U складається з усіх рядків досліджуваної таблиці.

Таблиця 2

Розподіл значень ознак між підгрупами, отриманих на останньому кроці агломеративної кластеризації

N=3532		N1=2960	N2=572	N=3532		N1=2960	N2=572
KHKS		0.25878	0.00175	PIK		0.47703	0.12587
	0	0.79349	0.20651		0	0.75586	0.24414
	1	0.99869	0.00130		1	0.95148	0.04852
AGE25		0	0.15734	GH		0.50270	0.20979
	0	0.85997	0.14004		0	0.76507	0.23493
	1	0	1		1	0.92537	0.07463
AGE20_40		0.00777	0.74301	REW		0	0.29021
	0	0.95234	0.04767		0	0.87938	0.12062
	1	0.05134	0.94866		1	0	1
AGE50		0.81622	0.17657	KV		0.18142	0.00699
	0	0.53596	0.46404		0	0.81009	0.18990
	1	0.95987	0.04013		1	0.99261	0.00739

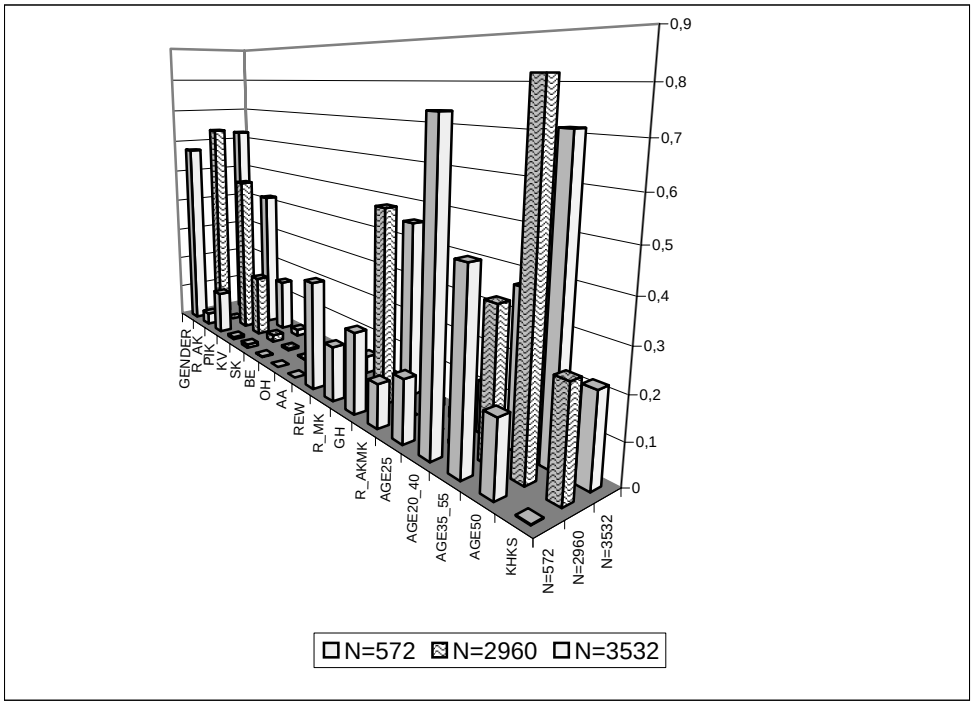

Рис. 3. Порівняльна характеристика значень ознак у групах з 572, 2960 та 3532 пацієнтів

Як було зазначено вище, на останньому кроці процесу кластеризації відбулося об’єднання в одну групу двох підгруп, в одній з яких виявились майже всі хворі із високими значеннями ознак PIK, KV, AGE50 та GH. Аналіз, що проведено із застосуванням кластерного аналізу, дозволив поділити всіх пацієнтів на групи, що характеризуються обмеженою кількістю атрибутів. Утворення групи з $k = 2$ показало, що в цю групу потрапили практично всі хворі.

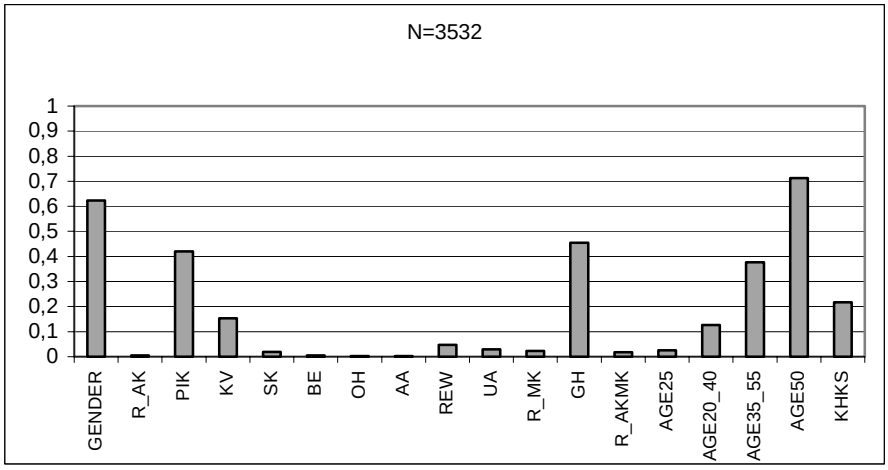

Рис. 4. Розподіл ознак у групі з 3532 пацієнтів

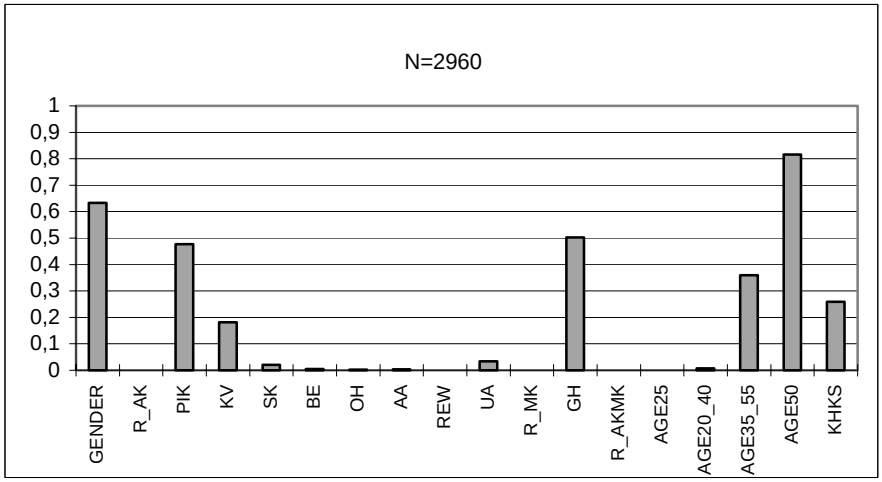

Рис. 5. Розподіл ознак у групі з 572 пацієнтів

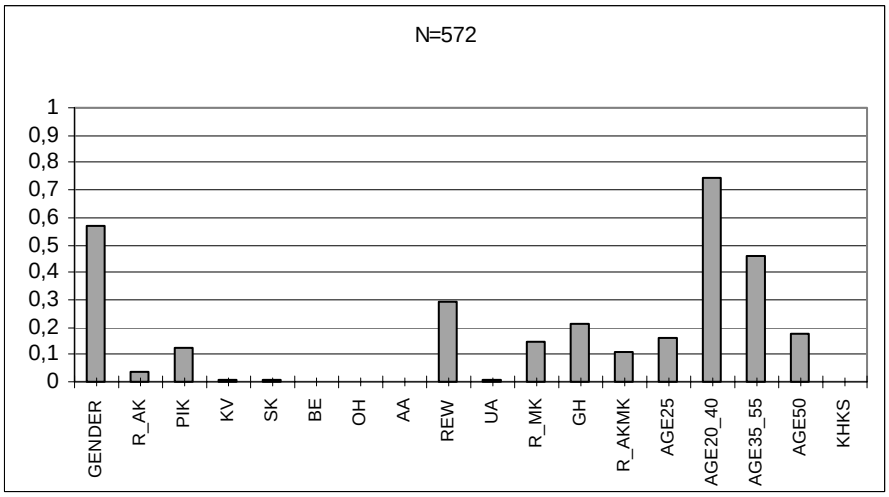

Рис. 6. Розподіл ознак у групі з 572 пацієнтів

Наявність атрибутів, значення яких зросло у цій групі, свідчить про можливий перерозподіл за групами і пацієнтів, що мають симптоми, які спричинили виникнення

вказаної хвороби. Надалі дані можна аналізувати у напрямку побудови правил прийняття рішень, які будуть використані для опису закономірностей утворення хвороби. У найзагальнішому випадку таке правило треба розглядати у вигляді:

$$(\mathrm{PIK}=1)\wedge(\mathrm{KV}=1)\wedge(\mathrm{AGE50}=1)\wedge(\mathrm{GH}=1)\Rightarrow \mathrm{KHKS}=1.$$

Таблиця 3

Правила формування груп пацієнтів за розподілом значень ознак на третьому етапі аналізу

1.	(99.2 %) IF AGE20_40= 0	TO GROUP 1 (95.2 %)
2.	(81.6 %) IF AGE50= 1	TO GROUP 1 (96.0 %)
3.	(50.3 %) IF GH= 1	TO GROUP 1 (92.5 %)
4.	(47.7 %) IF PIK= 1	TO GROUP 1 (95.1 %)
5.	(25.9 %) IF KHKS= 1	TO GROUP 1 (99.9 %)
6.	(18.1 %) IF KV= 1	TO GROUP 1 (99.3 %)
7.	(74.3 %) IF AGE20_40= 1	TO GROUP 2 (94.9 %)
8.	(29,0 %) IF REW= 1	TO GROUP 2 (100 %)
9.	(15.7 %) IF AGE25= 1	TO GROUP 2 (100 %)
10.	(14.5 %) IF R_MK= 1	TO GROUP 2 (100 %)

Вилученням з цього правила атомарних формул можна отримати інші правила для побудови процедур приняття рішень на групах пацієнтів, які утворювались в процесі ієрархічної кластеризації. Зокрема, можна припустити, що утворення нових груп при обході дерева, дендрограми згори донизу у вершинах з $k=7$ та $k=15$ приведе до необхідності перегляду правил через появу нових ознак, за якими будуть класифікуватись правила постановки діагнозу.

ВИСНОВКИ

Використання методу кластерного аналізу дозволило виділити у таблиці прийняття рішень групу ознак, що можуть бути використані як початкові для формування множини вирішальних правил в системі прийняття рішень щодо встановлення діагнозу. Аналіз реальної таблиці прийняття рішень показав, що виділення таких ознак дозволить скоротити загальний обсяг обчислень при формуванні системи правил та знаходження їх характеристик. Подальші дослідження дозволять оцінити якість правил, що можуть бути виведені із загального правила та зміну якості діагнозу, отриманого за допомогою моделі прийняття рішень.

Реалізація методу кластеризації та необхідні при цьому обчислення проведено М. Давидовим.

1. Margaret H.Dunham. Data Mining. Introductory and Advanced Topics. Pearson Education Inc, 2003. 2. Jiawei Han, Micheline Kamber. Data Mining:Concepts and Techniques. Morgan Kaufmann Publishers, 2001. 3. L.Kaufman, P.J.Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. New York: John Wiley & Sons,1990. 4. Годич О.В., Нікольський Ю.В., Щербина Ю.М.. Застосування штучної нейронної мережі типу SOM для розв'язування задачі діагностування // Вісн. Нац. ун-ту "Львівська політехніка". – 2002. – № 464. – С. 31–43. 5. Нікольський Ю.В., Щербина Ю.М., Якимечко Р.Я.. Дерева прийняття рішень та їх застосування для прогнозування діагнозу у медицині // Вісн. Львів. ун-ту. Сер. Прикл. мат

та інформа. – 2003. – Вип. 4. – С. 191–211. 6. Jean-Marc Adamo. Data Mining for Association Rules and Sequential Patterns: Sequential and Parallel algorithms. Springer-Verlag. 2000. 7. Szczuka M. Rules as attributes in classifier construction // Proceedings of RSFDGrC'99, Yamaguchi, Japan, LNAI 1711, Springer-Verlag, Berlin. – P. 492–499. 8. Slowinski K., Stefanowski J. On limitations of using rough set approach to analyse non-trivial medical information systems // Proceedings of the 4th Int. Workshop on Rough Sets, Fuzzy Sets and Machine Discovery, Tokyo 1996. – P. 176–184. 9. Richard O. Duda. Pattern Classification. John Wiley&Sons, Inc. 2000. 10. Christopher M. Bishop. Neural Networks for Pattern Recognition.Oxford University Press. 2003. 11. Komorowski J., Pawlak Z., Polkowski L., Skowron A. Rough Sets: A Tutorial // A New Trends in Decision Making, Springer-Verlag, 1999. – P. 3–98.

УДК 004.652.4+004.852

Ю.М. Оградіна, П.Ф. Павлов,
Р.В. Разуваєв, В.В. Шепелєв, П.І. Кузьменко
Харківський національний університет радіоелектроніки,
кафедра "Соціальна інформатика"

РОЗРОБКА СИСТЕМИ АНАЛІЗУ НАПІВСТРУКТУРОВАНИХ ІНФОРМАЦІЙНИХ ПОВІДОМЛЕНЬ

Paper describes the program system for analysis of natural language messages about emergencies. Using data mining methods (association rules) for analyzing data are proposed.

Описано програмну систему аналізу повідомлень природною мовою про надзвичайні ситуації. Для проведення аналізу пропонується використовувати методи інтелектуального аналізу даних, зокрема побудову асоціативних правил

На початку XXI століття, з різким, «вибуховим» збільшенням розмірів баз даних та інформаційних потоків зросла актуальність ефективного використання даних. Без розвитку засобів і методів опрацювання інформації велика її частка може стати "мертвим вантажем", що "засмічує" канали передачі і носії.

До початку 90-х років для цих цілей застосовувалася, в основному, прикладна статистика. Однак її методи слабко пристосовані до реальних задач, оскільки оперують умовними "усередненими" величинами. Крім того, традиційні статистичні методи в силу своєї числової природи не можуть дати інструмента ані для аналізу текстової інформації, ані для пошуку складних, неочевидних залежностей між різними даними [1]. Виникнення інтелектуального аналізу даних (ІАД, Data Mining) по праву вважається продуктом природної еволюції інформаційних технологій в останні роки [2].

У роботі досліджується задача опрацювання та аналізу повідомлень про надзвичайні ситуації (НС), що були зібрані Міністерством Російської Федерації в справах Цивільної оборони, надзвичайних ситуацій та ліквідації наслідків стихійних лих (МНС РФ). Ці повідомлення складають великий обсяг слабко структурованої інформації, що викладена