УДК 004.9

Коробчинський М. В.[1], Чирун Л. Б.[2], Висоцька В. А.[3], Нич М. О.[4]

[1]Д-р техн. наук, старший науковий співробітник Військово-дипломатичної академії імені Євгена Березняка, Київ, Україна
[2]Провідний спеціаліст інституту комп'ютерних наук та інформаційних технологій Національного університету «Львівська політехніка», Львів, Україна
[3]Канд. техн. наук, доцент, доцент кафедри «Інформаційні системи та мережі» Національного університету «Львівська політехніка», Львів, Україна
[4]Магістр кафедри «Інформаційні системи та мережі» Національного університету «Львівська політехніка», Львів, Україна

ОСОБЛИВОСТІ ПРОГНОЗУВАННЯ РЕЗУЛЬТАТІВ МАТЧІВ У КІБЕРСПОРТІ

Актуальність. Зараз є актуальним розроблення систем прогнозування матчів у кіберспоті, це пов'язано з активним розвитком кіберспорту. У цій статті реалізовано можливість прогнозувати матчі користувачів.

Мета. Метою виконання роботи є проектування моделі системи колективного прогнозування результатів ігор у кібер-спорті з використанням сучасної технології нейропрогнозування. Завданням є розроблення системи для спільного користувацького прогнозування результатів кіберспортивних матчів та самостійного опрацювання інформації і видачі власного прогнозу. До основних задач належать наступні: облік та аналіз всіх минулих і майбутніх ігор; облік та аналіз характеристик/результатів всіх команд; надання можливості користувачу персонально робити прогноз на кожен матч; визначення шансу на виграш команди на основі даних за попередні матчі.

Метод. Проблему вирішено методом опитування експертів шляхом проведення аналітичних записок та за допомогою штучної нейронної мережі. В створеній нейромережі є три шари. Перший шар складається із 10 нейронів – рецепторів, або нейронів вхідних даних. Другий шар нейронів є внутрішній. Третій шар нейронів є вихідний, в ньому є лише 2 нейрони. Вхідною інформацією для алгоритму є кількість виграних матчів з останніх 10; кількість виграних матчів перед даною зустріччю (вінстрік); рейтинг команди; стабільність складу (час незмінності складу команди); середній показник програшів даної команди. Відповіддю є 1 або 2 (перемога конкретної команди).

Результати. Для досягнення результату проведений аналіз відповідної літератури з інформацією про основні види колективного прогнозування. Розроблено дерево мети, і проведено систематичний аналіз предметної області. Застосовано метод інтерв'ю з експертами. Інтернет-ресурси реалізовані за CMS Drupal. Проаналізовані основні методи колективного прогнозування. Проведений систематичний аналіз об'єкта дослідження і предмета, цілей, побудованих дерев, визначено проблему і побудовано UML-діаграми. Проаналізовано застосування методу інтерв'ю з експертами. Реалізовано веб-сайт з CMS Drupal і мови програмування PHP.

Висновки. На основі розробленого алгоритму розрахунку прогнозів та навчання нейромереж реалізовано незалежний від людського фактору процес прогнозування матчів в кібер-спорті. Наявність такої системи значно спростить пошук прогнозів на кібер-спортивні матчі та дасть можливість кожному бажаючому прийняти участь у прогнозуванні матчів. Система дає новий поштовх до вирішення проблеми прогнозування результатів не лише у кібер-спорті, а і у спорті взагалі.

Ключові слова: кіберспорт, інформаційна система.

НОМЕНКЛАТУРА

ІС – інформаційна система;

ПО – предметна область;

ШНМ – штучні нейронні мережі;

БНМ – біологічних нейронних мереж;

ККД – коефіцієнт корисної дії;

S – система прогнозування кібер-матчів;

X – вхідні дані ІС з різних достовірних джерел-сайтів (наприклад, історія ігор команд та учасників команд в інших змаганнях);

C – внутрішній контент системи, набутий під час її функціонування;

N – штучна нейронна мережа;

T – часовий проміжок аналізу діяльності команди;

E – експертна оцінка;

Q – запити від користувачів ІС;

U – обмеження на аналіз даних для прогнозування;

Y – вихідні дані ІС як результат прогнозування;

ϕ – функція збору даних для прогнозування;

φ – функція навчання ШНМ;

θ – функція прогнозування результатів матчів;

$N_п$ – кількість користувачів, що прийняли участь у голосуванні;

K_i – кількість користувачів, що проголосували за i-ту команду.

ВСТУП

Відеоігри з'явилися нещодавно та темпи їх еволюції вражає: від простеньких дитячих іграшок на автоматах з піксельною графікою до комп'ютерних ігри у вигляді інтерактивного кіно (спецефекти, озвучки від акторів і моделі для лицьової анімації). Еволюція ігор відбулась як через технологічні вдосконалення, так завдяки появі Інтернет вони стали багатокористувацькими. Вперше кіберспортивні риси та змагання по мережі з'явилися в Quake: Arena, Starcraft. Ігри захоплювали не графікою, а відмінним динамічним геймплеєм, вимагали від гравців нелюдською реакції і багатьох годин тренувань. Online змаганнями не обмежувалися, люди почали збиратися на LAN-турніри в Інтернет-клубах [1]. Активно кіберпорт почав розвиватись на протязі останніх 5–10 рр із-за доступності Інтернет.

Стрімкий розвиток кіберспорту значно вплинув на ігрову індустрію [1]. Розробники випускають ігри, в яких гравці об'єднуються в команди і змагаються між собою.

DOI 10.15588/1607-3274-2017-3-11

Виникли турніри, професійні команди, і фанати. Виникла потреба прогнозування переможця. Зазвичай процес прогнозування покладаний на вузьке коло так званих експертів з цього питання – букмекерів або людй, які постійно спостерагають за ходом гри на конкретному інформаційному ресурсі. На ІС інформаційного ресурсу кіберспорту покладені лише функції фіксації статистики матчів та ігроків. Це унеможливлює взяти участь в успішному прогнозуванні результатів матчівдля простих спостерігачів змагань в кіберспорті. Метою виконання роботи є проектування моделі системи колективного прогнозування результатів ігор у кібер-спорті з використанням сучасної технології нейропрогнозування.

1 ПОСТАНОВКА ЗАДАЧІ

Застосування методу нейромережного прогнозування на основі аналізу статистики попередніх результатів матчів та рівня професійності команд та ігроків значно спрощує процес отримання успішного прогнозу для простих спостерігачів змагань в кіберспорті. Завданням є розроблення загальної архітектури ІС для спільного користувацького прогнозування результатів кіберспортивних матчів та самостійного опрацювання інформації і видачі власного прогнозу. ІС має навчатися на сонові зібраної статистики змагань на протязі певного періоду часу з врахуванням учасників змагань та їх профілів.

Формальною моделлю системи прогнозування кібер-матчів є кортеж

$$S =< X, C, N, T, E, Q, U, Y, \phi, \varphi, \theta > .$$

До основних задач ІС належать наступні: облік та аналіз всіх минулих/майбутніх ігор; облік та аналіз характеристик/результатів всіх команд; надання можливості користувачу персонально робити прогноз на кожен матч; визначення шансу на виграш команди на основі даних за попередні матчі. Якщо з першими трьома задачами успішно справляються більшість ІС прогнозування матчів у кіберспоті, то з останньою задачею виникають певні складнощі із-за того, що простий користувач ІС має залежити від думки еспертів – окремих букмекерів, тобто від людського фактору. Автоматичне прогнозування результатів матчів та можливість системи начатися на протязі певного періоду часу сприятиме покращенню результатів прогнозування та вносить елемент незалежності прогнозування від людини.

Вхідні дані з різних джерел X (стабільність складу команди), внутрішній контент ІС C (наприклад, історія ігор на цьому сайті, зміна учасників команди, активність команди тощо), експертні оцінки E (рейтинг командб середній показник програшів даної команди) та обмеження на прогнозування U (наприклад, аналіз лише останніх 10 матчів) за певний проміжок часу T (наприклад, матчі лише за останні рік або час незмінності складу команди) формують підґрунтя для процесу прогнозування штучною нейронною мережею N

$$Y = \theta \circ \varphi \circ \phi,$$

де $C = \phi(X, T, E, U)$, $N = \varphi(C, T, U)$ та $Y = \theta(Q, N, T, U)$. Відповіддю є 1 або 2 (перемога конкретної команди).

2 ОГЛЯД ЛІТЕРАТУРИ

Методи прогнозування – це різноманітні прийоми і способи мислення, що дають змогу на основі аналізу ретроспективних даних, екзогенних і ендогенних зв'язків об'єкта прогнозування, а також їхніх змін у межах розглянутого явища і процесу, вивести твердження певної вірогідності відносно майбутнього розвитку об'єкта [2]. Відомо більше за 150 методів і прийомів у прогнозуванні. Кожен метод має особливості залежно від мети його використання і рівня проведених досліджень. Вибір методів прогнозування здійснюється відповідно до характеру об'єкта і вимог, які висуваються до інформаційного забезпечення [2–5]. Для спільного користувацького прогнозування результатів кіберспортивних матчів достатньо використовувати інтуїтивні методи прогнозування. Вони дають змогу отримати прогнозну оцінку стану розвитку об'єкта в майбутньому незалежно від інформаційної забезпеченості. Будується раціональна процедура інтуїтивно-логічного мислення людини в поєднанні з кількісними методами оцінки й опрацювання отриманих результатів. Узагальнена думка експертів та їх експертиза є вирішенням проблеми [2].

Процес експертної оцінки включає такі напрями: формування експертної групи; підготовку і проведення експертизи; статистичне опрацювання отриманих результатів опитування [3, 5]. Етапи проведення експертних оцінок: постановка проблеми; відбір експерта; опитування експерта; опрацювання експертних оцінок. Методи експертних оцінок поділяють: за кількістю експертів: індивідуальні та колективні; за технологією опрацювання інформації: прямі і експертні методи зі зворотним зв'язком; за технологією отримання прогнозної оцінки [5]. Процес експертної оцінки краще реалізувати через штучні нейронні мережі. Це математичні моделі, а також їх програмні або апаратні реалізації, побудовані за принципом організації й функціонування біологічних нейронних мереж – мереж нервових кліток живого організму. ШНМ не програмують у звичному розумінні цього слова, вони навчаються. Можливість навчання – одне з головних переваг нейронних мереж перед традиційними алгоритмами [6–7]. Задачі, які вирішують ШНМ, зводяться до апроксимації багатовимірних функцій, тобто побудови відображення $F : x \to y$ [7]. Якщо розглядати ШНМ як деяке середовище для опрацювання інформації, тоді її можна задати шляхом визначення елементів даного середовища та правил їх взаємодії. В цьому випадку ШНМ є структурою, яка складається з великої кількості процесорних елементів, кожен з яких має локальну пам'ять і може взаємодіяти з іншими процесорними елементами за допомогою комунікаційних каналів з метою передачі даних, що можуть бути інтерпретовані довільним чином. Процесорні елементи незалежно в часі опрацьовують локальні дані, що поступають до них через вхідні канали. Зміна параметрів алгоритмів такої опрацювання залежить тільки від характеристик даних. ШНМ – обчислювальні парадигми, які реалізують спрощені моделі БНМ (локальні ансамблі нейронів, об'єднані синаптичними зв'язками; сукупність ансамблів формує мозок із різноманітними функціональних можливостями).

3 МАТЕРІАЛИ І МЕТОДИ

Оскільки на даний момент кіберспорт почав швидко розвиватись, то зявилась потреба в ресурсі, де фанати зможуть робити прогнози майбутніх матчів та дізнатись результати минулих. Наявність ІС колективного прогнозування результатів ігор у кібер-спорті дає можливість користувачеві отримати доступ до необхідної інформації використовуючи один ресурс. Користувач може переглянути результати минулих матчів, подивитись оцінку шансів на виграш кожної з команд на майбутніх матчах та самому зробити прогноз на ту чи іншу команду. Призначенням такої системи є збір оціночних тверджень користувачів відносно кожного матчу та інтелектуальне прогнозування результату даного матчу, також система порівнює прогноз із результатом і може визначити відсоток матчів, які були спрогнозовані користувачами правильно. ІС колективного прогнозування результатів ігор у кібер-спорті повинна розв'язувати такі задачі: реалізація методу прогнозування на основі збору голосів користувачів; визначення ефективності методу (порівняння прогнозу з реальними результатами); швидкий та зручний пошук інформації (результатів попередніх ігор).

Необхідність створення такої системи викликана відсутністю на даний момент реалізованого аналогу, який би об'єднував в собі результати і прогнози не з якоїсь певної гри, а з декількох; а також тим, що аналоги є платними. Інтуїтивні методи прогнозування – це методи вирішення проблеми за допомогою інтуїції, досвіду та думки експертів-учасників даних методів. Метод експертних оцінок оснований на проведенні інтуїтивного та логічного аналізу проблеми з кількісною оцінкою і опрацюванням результатів. Після цього узагальнена думка експертів вважається рішенням даної проблеми. Використання інтуїції, логічного мислення та кількісних оцінок з формального опрацювання дозволяє одержати ефективне вирішення проблеми. Особливостями методу експертних оцінок є, по-перше, обгрунтована організація проведення всіх етапів експертизи забезпечує найбільшу ефективність роботи на кожному з етапів; по-друге, застосування кількісних методів як при організації експертизи, так і при оцінці суджень експертів і формальній груповіму опрацюванні результатів. Найбільш часто ці методи використовують при розгляді соціально-економічних проблем, де неможливо виробити формалізовану прогностичну модель [4–5, 8–10]. Метод експертних оцінок проводять трьома методами як інтерв'ю, аналітичні записки або написання сценарію. Проблему вирішено методом опитування експертів шляхом проведення аналітичних записок та за допомогою ШНМ. Етапи застосування методу колективних експертних оцінок:

– формування експертної групи (відбір експертів проводиться з використанням методів самооцінки, методи взаємної оцінки, по минулому досвіду);

– визначення компетентності експертів;

– оцінка показності або репрезентативності групи;

– отримання індивідуальних суджень експертів по заданій проблемі;

– узагальнення думок про відносну важливість заданої проблеми експертами;

– оцінка ступеня узгодженості експертів з урахуванням коефіцієнта варіації;

– побудова гістограми розподілу думок експертів;

– формулювання плану прогнозу.

Технологія узагальнення оцінок залежить від виду оцінок, кількості даних оцінок. Якщо прогнозом є число, то може використовуватися середнє значення. Якщо відповіді експертів різні, то використовується теорія ігор. Якщо оцінка експертів давалася у вигляді рангів, то в процесі узагальнення визначається сума рангів й узгодженість оцінок за допомогою коефіцієнта конкордації; Спірмана або Кендела. Про достовірність групових експертних оцінок зазвичай судять по їх узгодженості. При проведенні експертних опитувань, як правило, отримують оцінки декількох об'єктів. Визначити узгодженість оцінок, які даються різними експертами, можна за допомогою непараметричного двохфакторного дисперсійного аналізу. При виконанні аналізу, в якості першого чинника розглядаються експерти, в якості іншого чинника – об'єкти, які оцінюються експертами. Рівні першого фактора – це різні експерти, а рівні другого чинника – різні об'єкти.

Узгодженість оцінок експертів визначається за відсутності впливу фактора, пов'язаного з експертами. У поширених статистичних пакетах для цього використовують критерій Фрідмана (Friedman) і, якщо є можливість ранжувати експертів за величиною оцінок, то критерій Пейджа (Page). Зазвичай, тестується гіпотеза «є відмінності між середніми значеннями оцінок деяких експертів» з оцінкою рівня значущості гіпотези. Якщо рівень значимості гіпотези не перевищує 5 або 10%, то можна вважати, що оцінки експертів узгоджені і достовірні. Колективна думка експертної групи може бути виражене у формах: кількісних оцінок у фізичних одиницях виміру або у вигляді відношення; бальних оцінок; попарних порівнянь; угруповань (сортування) [5]. Реалізація методу опитування експертів полягає в голосуванні кожного користувача (який вважається експертом) із $N_п$.

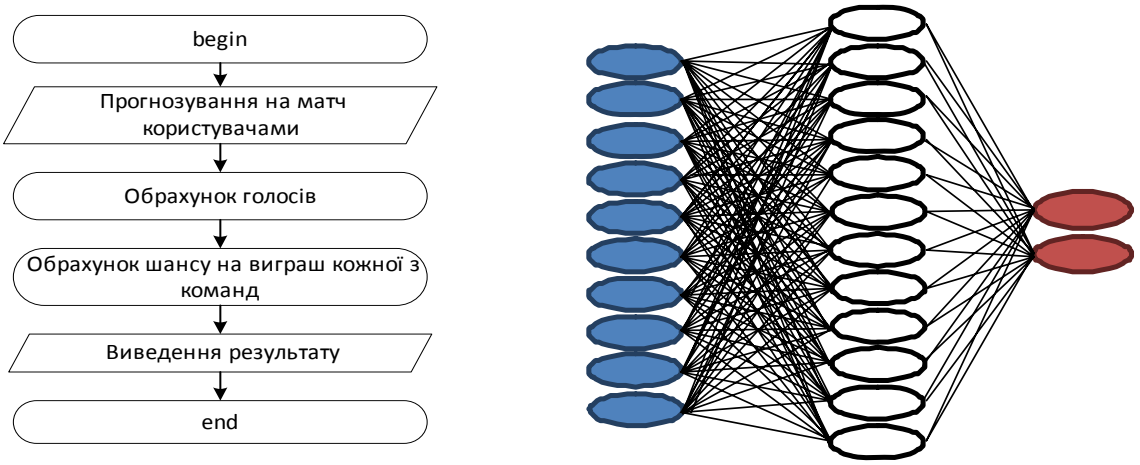
Алгоритм працює наступним чином (рис. 1 а):

Крок 1. Всі бажаючі користувачі із $N_п$ роблять прогноз на даний матч.

Крок 2. Розрахунок сумарної кількісті голосів та за кожну з команд з врахуванням K_i.

Крок 3. На основі голосів здійснюють прогнозування шансу на перемогу кожної з команд у відсотках. Результатом є шанс на перемогу кожної з команд у даному матчі.

В створеній ШНМ є три шари (рис. 1б). Перший шар складається із 10 нейронів – рецепторів, або нейронів вхідних даних. Другий шар нейронів є внутрішній. Третій шар нейронів є вихідний, в ньому є лише 2 нейрони. Вхідною інформацією для алгоритму є кількість виграних матчів з останніх 10; кількість виграних матчів перед даною зустріччю (вінстрік); рейтинг команди; стабільність складу (час незмінності складу команди); середній показник програшів даної команди. Відповідью є 1 або 2 (перемога конкретної команди).



а б

Рисунок 1 – Прогнозування:
а – алгоритм обрахунку прогнозів; б – схема нейронної мережі

4 ЕКСПЕРИМЕНТИ

Для створення ШНМ використано бібліотеку Fast Artificial Neural Network Library (FANN). Діаграма варіантів використання описує функціональне застосування ІС колективного прогнозування матчів у кіберспорті (рис. 2) [1, 8–10].

Користувачу для прогнозування потрібно пройти автентифікацію. Тоді він може переглянути результати попередніх ігор, які його цікавлять. Користувач може передивитись результати прогнозів інших користувачів на даний матч. Після прогнозування результат заноситься в базу даних і об'єднується з іншими прогнозами на даний матч (рис. 3а).

Для отримання результату обрахунку прогнозу користувач повинен завантажити ІС з веб-сайту, ввести потрібні дані про команди, після чого програма видає

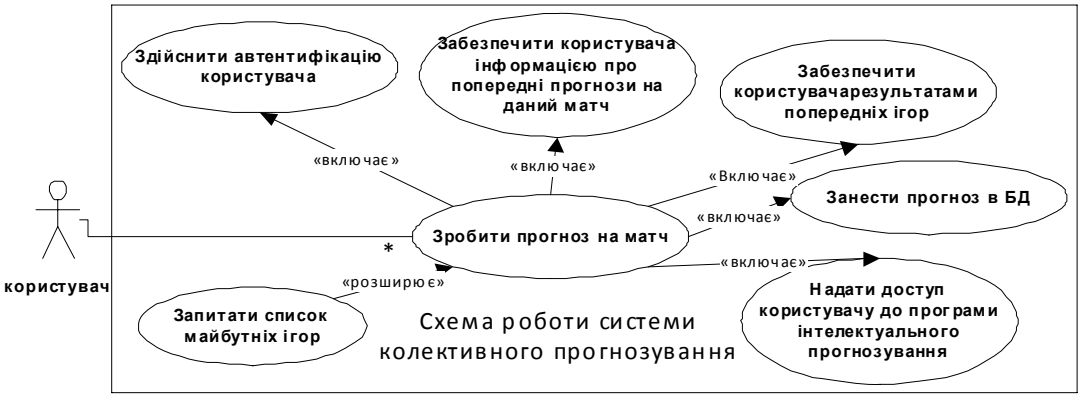
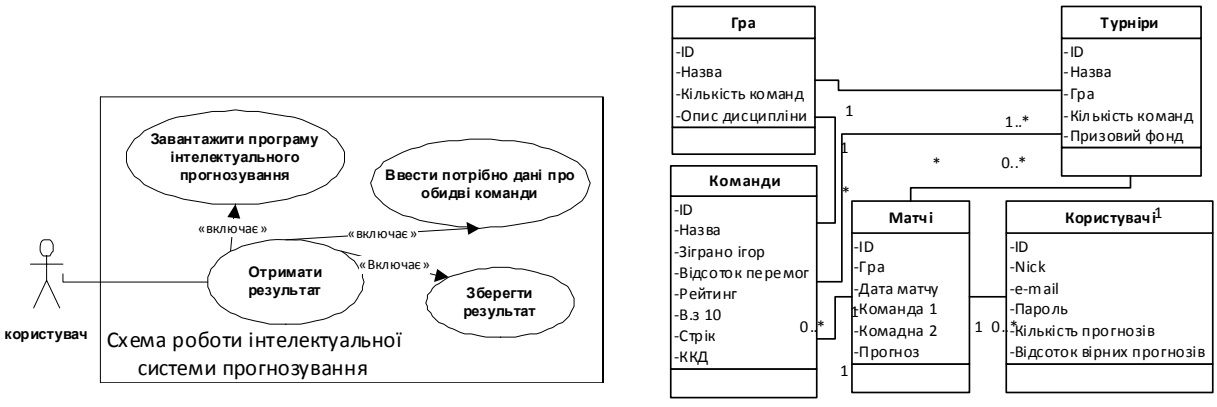


Рисунок 2 – Діаграма варіантів використання ІС прогнозування матчів у кіберспорті



а б

Рисунок 3 – Діаграма ІС:
а – варіантів використання; б – класів

результат, який можна зберегти в текстовий файл. Діаграма класів ІС прогнозу результатів матчів наведена на рис. 3б. Діаграма на рис. 4а демонструє процес роботи ІС починаючи від перевірки наявності доступу до сайту, де розміщена система, до виводу результату користувачеві.

– З'єднання з сайтом. Користувач повинен залогінитись на сайті; якщо він ще не зареєстрований, то він проходить реєстрацію.

– Користувач має обрати одну з 3-х вкладок – прогнозування майбутніх матчів, перегляд результатів минулих або завантажити програму прогнозування.

– Якщо користувач завантажив програму інтелектуального прогнозування, то йому потрібно ввести вхідні дані обох команд.

– Якщо користувач вибрав один з інших варіантів, то він обирає дисципліну, в рамках якої проходить матч, що його цікавить.

– Потім користувач повинен обрати турнір, в рамках якого проходить матч.

– Якщо користувач обрав прогноз, то після вибору потрібного йому матчу він може зробити свій прогноз на даний матч.

– В кінцевому результаті користувач отримує результат прогнозування свого, колективного або інтелектуального.

ІС складається з наступних частин: команди (рис. 4б), опитування (рис. 4в), ігри, програма інтелектуального прогнозування. Вид сторінок опитування включає в себе сторінки, на яких власне проводиться прогнозування результатів матчів, для реалізації цього компоненту в системі використовується модуль Drupal Advanced Poll.

Структура сутності «Опитування» складається з таких полів (рис. 4а): Poll name – власне назва сторінки з даним опитуванням; гра – гра (дисципліна) з якої проводиться даний матч; дата проведення матчу; опис – короткий опис матчу, турнір в рамках кого проводиться матч, стадія; команда 1; команда 2. Вид сторінки команди включає загальну сторінку з переліком команд, та персональну сторінку кожної команди з детальної інформацією про неї. Адміністратор може в будь-який час додати нову команду, якщо в цьому є потреба.

На рис. 5 подана діаграма компонентів ІС прогнозування результатів ігор у кіберспорті. ІС працює з базою даних за допомогою інтерфейсу «IDialog». ІС використовує модулі, такі як User_Api, Advanced Poll,Viev reference, CK_Editor. ІС використовує три бібліотеки: fann (Fast Artificial Neural Network Library), fannj (додаткова бібліотка для роботи з fann), та бібліотеку jna (Java Native Acces).

Вид сторінок «ігри» включає в себе 3 сторінки, на яких знаходяться всі прогнози по минулих чи майбутніх матчах з трьох дисциплін (ігор), та короткий опис даної гри. На рис. 6а наведено приклад однієї з таких сторінок. ІС створена на основі ШНМ і «навчена» за допомогою даних про попередні матчі між командами. Для створення ШНМ завантажують три бібліотеки: libfan, fannj та jna. З таблиці навчання створеної в третьому розділі створюється файл, який буде мати в собі набір тих самих «уроків» (рис 6б). Тепер створимо ШНМ, «навчимо» її, та збережемо в файл (рис. 7). ШНМ складається із шарів (layer) нейронів, в першому шарі у нас 10 нейронів: виграно з 10 останніх матчів першої команди; виграшний

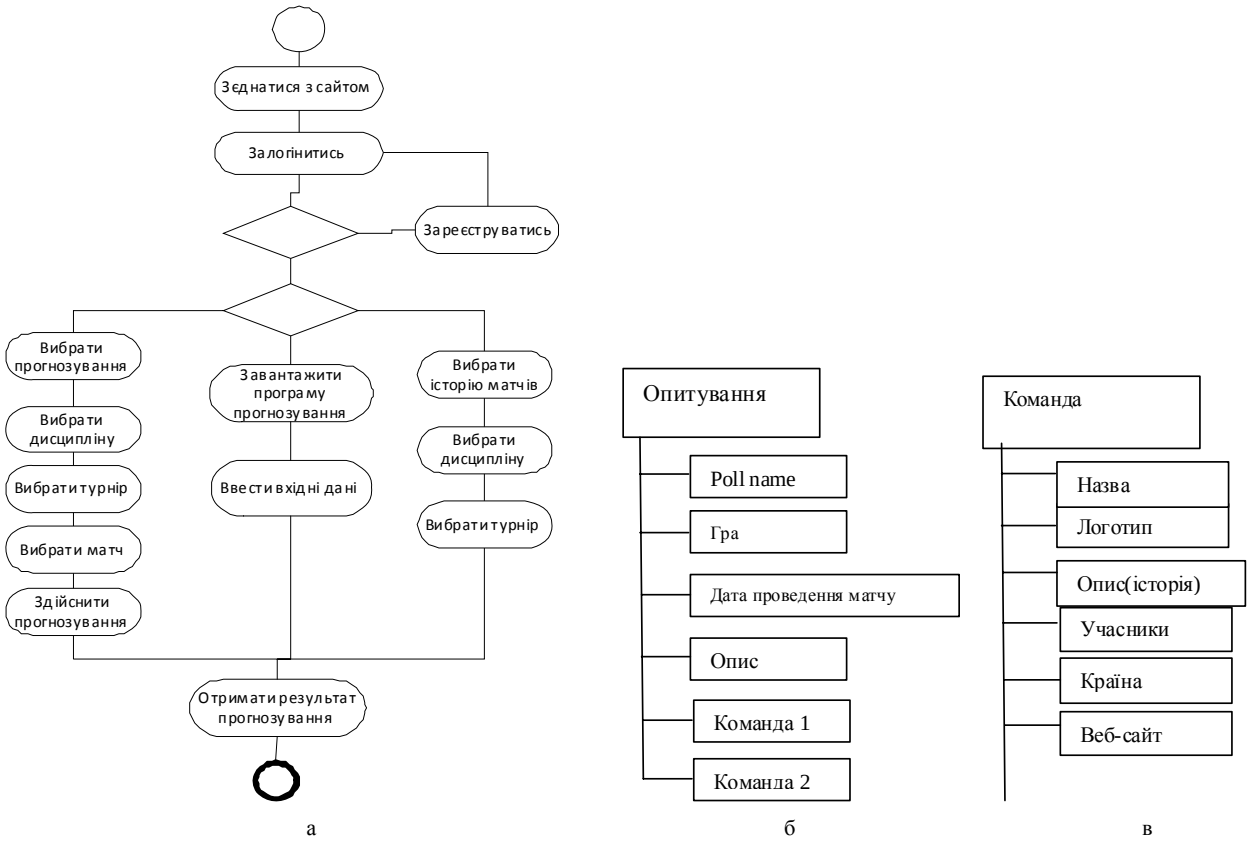


Рисунок 4 – Структура сутності ІС:
а – діаграма діяльності; б – структура сутності «Опитування»; та в – «Команда»

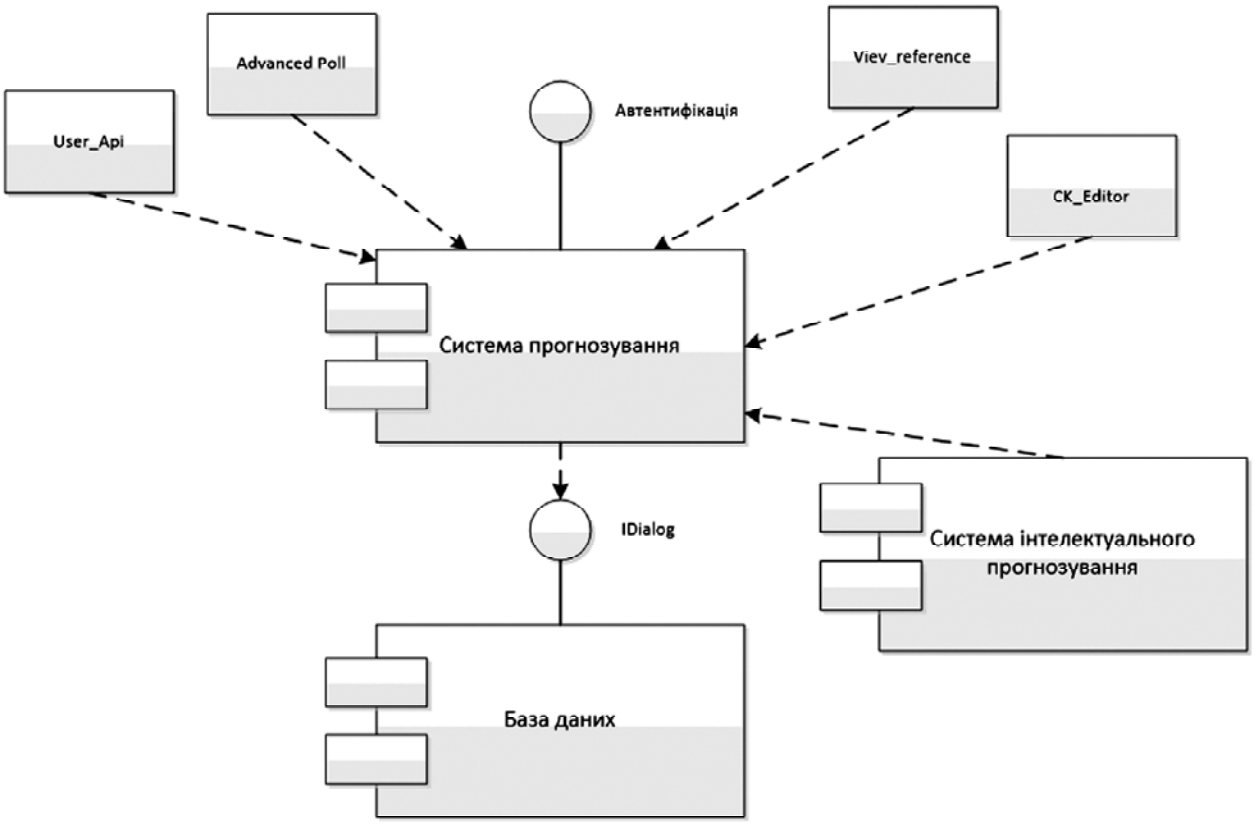
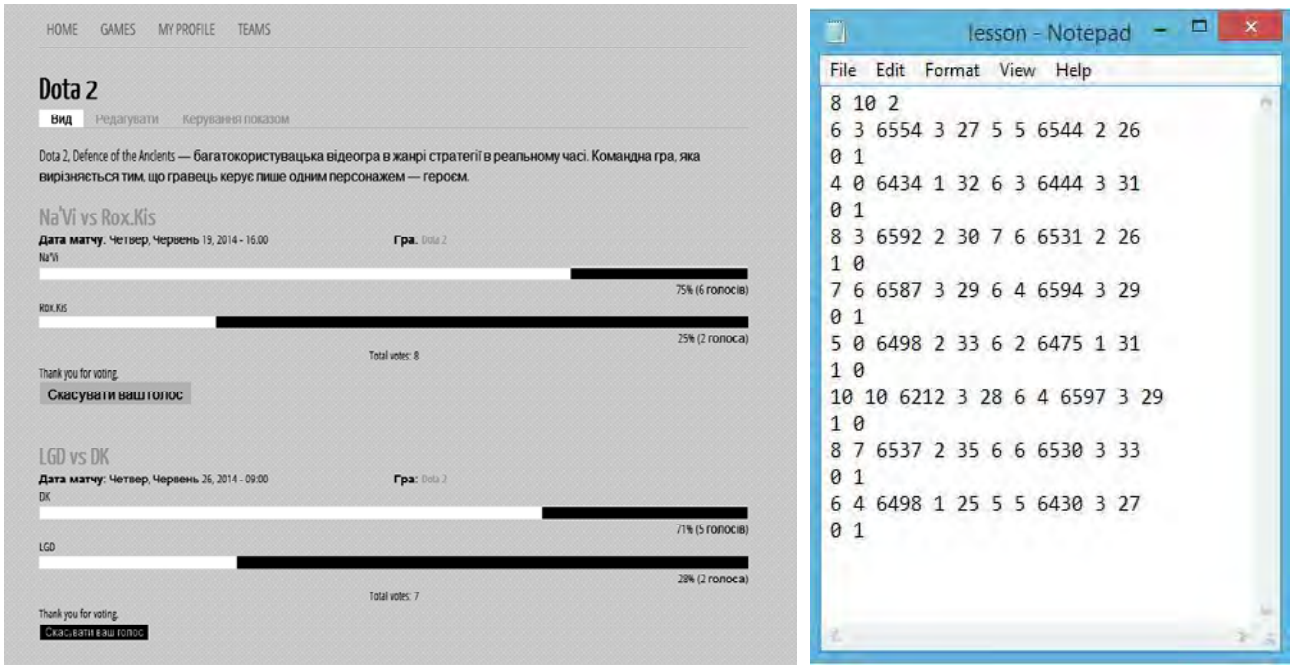


Рисунок 5 – Діаграма компонентів



а б

Рисунок 6 – Дані для навчання:
а – сторінка гри з прогнозами на матчі; б – набір «уроків» для ШНМ

стрік першої команди; рейтинг першої команди; стабільність складу першої команди; ККД першої команди (на прикладі атакуючого удару); виграно з 10 останніх матчів другої команди; виграшний стрік другої команди; рейтинг другої команди; стабільність другої першої команди; ККД другої команди.

В останньому шарі маємо два нейрони, які відповідають першій та другій команді. Об'єкт класу fann – це і є мережа, яка створюється на основі раніше створених шарів. Об'єкт класу trainer інкапсулює алгоритми навчання ШНМ переданої при створенні тренера, після навчання зберігається у файл. Першим етапом реалізації про-

екту є проектування бази даних (БД). Всі сутності будуть подані таблицями в базі даних. Призначення відношень бази даних:

– Гра – Ігри (дисципліни) з якими працює система.

– Атрибути : _id, Назва, Кількість_Команд, Опис;

– Матчі – всі минулі і майбутні матчі занесені в систему.

– Атрибути:_id, Poll_Name, ID_Гри, Дата,Команда_1,Команда_2,Прогноз;

– Команди – інформація про конкретну команду.

– Атрибути:_id, Назва, Зіграно_ігор, вінрейт, рейтинг, стрік, ккд,;

– Користувачі – інформація про користувачів.

– Атрибути:_id, Нікнейм, e-mail, Пароль, Кількість_прогнозів, Відсоток_вірних_прогнозів;

– Турніри:_id, Назва, Гра, Кількість_команд, призові.

Система розпізнає 3 види користувачів: незареєстрований користувач, зареєстрований користувач та адміністратор. Незареєстрований користувач вільно пересувається сайтом, проте він не бере участі в прогнозуванні, лише переглядає результати. Зареєстрований користувач також вільно пересувається сайтом, переглядає інформацію про команди, їх попередні та майбутні матчі, приймає участь у прогнозуванні матчів та має доступ до прогнозування. Адміністратор має всі права доступу на сайті, його задача оновлювати дані про команди та додавати нові матчі для прогнозування. Інтерфейс системи є графічний. На кожній зі сторінок користувач в верхньому меню має 4 кнопки: HOME – перехід на головну сторінку; GAMES – перехід на сторінку зі всіма матчами незалежно від дисципліни; MY PROFILE – перехід на сторінку профілю користувача; TEAMS – перехід на сторінку списку команд. На головній сторінці користувачу доступне інтерактивне меню з трьох слайдів: кожен слайд відповідає за одну дисципліну, вони перегортаються автоматично кожні 4 секунди, та користувач при бажанні може перейти на потрібний слайд за допомогою кнопки внизу. На кожному зі слайдів є посилання на сторінку кожної з дисциплін (ігор), де розміщено список всіх матчів з даної дисципліни. На сторінці гри користувач бачить список всіх матчів, і результатів їх прогнозування, в разі якщо користувач не спрогнозував результат матчу, йому пропонується спрогнозувати його (рис. 6а).

Внизу кожної сторінки присутні три форми, в яких користувач може дізнатись про останні додані на сайт матчі, рейтинг команд та перейти на випадковий матч і переглянути результат його прогнозування, або спрогнозувати його, якщо ще не зробив (рис. 8а).

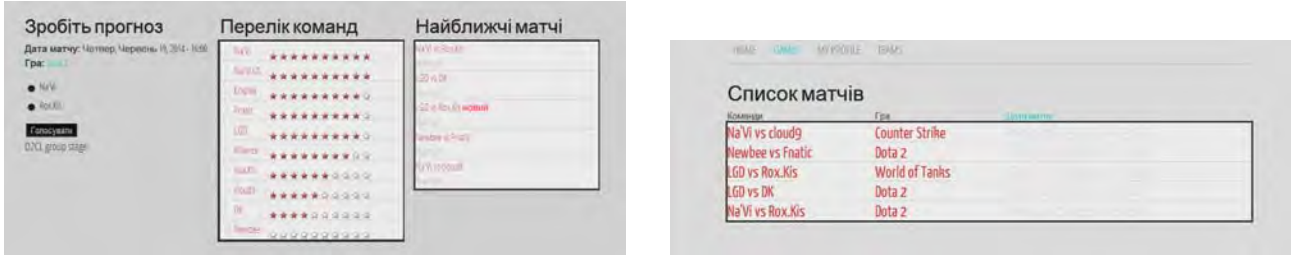
Сторінка Games містить останні додані матчі незалежно від дисципліни (рис. 8б), користувач може перейти з цієї сторінки на сторінку матчу (рис. 6а), який його цікавить і проголосувати, чи подивитись результати голосування.

5 РЕЗУЛЬТАТИ

ІС прогнозування результатів матчів (рис. 9) має досить простий інтерфейс: користувачеві достатньо ввести дані обох команд і натиснути кнопку, після чого програма показує, яка команда на її думку переможе. Перевірка показала що програма передбачає правильно близько 70% матчів (13 із 19). Вхідними даними є: голоси, які користувач віддає певним командам по кожному з матчів; дані команд суперників. Результатами роботи системи є: шанс перемоги кожної з команд, по даному матчу обрахований коллективно; переможець зустрічі визначений інтелектуальною складовою системи.

```
public static void main(String[] args){
 //Спочатку створюєм шари
List<Layer> layerList = new ArrayList<Layer>();
layerList.add(Layer.create(10, ActivationFunction.FANN_SIGMOID_SYMMETRIC, 0.01f));
layerList.add(Layer.create(13, ActivationFunction.FANN_SIGMOID_SYMMETRIC, 0.01f));
layerList.add(Layer.create(2, ActivationFunction.FANN_SIGMOID_SYMMETRIC, 0.01f));
Fann fann = new Fann(layerList);
 //Створюєм тренера і визначаємо алгоритм навчання
Trainer trainer = new Trainer(fann);
trainer.setTrainingAlgorithm(TrainingAlgorithm.FANN_TRAIN_RPROP);
      /* Проводимо навчання мережі, з максимальною кількістю циклів 100000, показуєм звіт
      кожну  100у ітерацію і досягаєм похибку менше 0.0001 */
      trainer.train(new File("lesson.data").getAbsolutePath(), 100000, 100, 0.0001f);
      fann.save("ann"); }
```

Рисунок 7 – Створення та навчання ШНМ



а б

Рисунок 8 – Форма:
а – внизу кожної зі сторінок; б – сторінки всіх матчів

Наприклад, користувач, який хоче спрогнозувати матч з дисципліни Dota 2, переходить на сторінку прогнозування (рис. 10а). Перебуваючи на даній сторінці користувач робить свої прогнози відносно майбутніх матчів, також він може перейти на персональну сторінку кожного з матчів (рис. 10б). Під кнопкою проголосувати користувач може побачити короткий опис матчу, тобто турнір в рамках якого проводиться даний матч, та стадія турніру. Вигляд сторінки після голосування (рис 10в).

Якщо користувач бажає отримати прогноз за допомогою програми інтелектуального прогнозування, він повинен завантажити програму з головної сторінки сайту, запустити програму і ввести дані обох команд. Після вводу користувач натискає кнопку «Порахувати» і отримує результат яка ж команда повинна перемогти (рис. 9). Створена ІС на основі отриманих даних вирішує, яка з команд переможе у матчі. Для цього використано ШНМ (табл. 1). ІС відомі такі дані: кількість перемог з останніх 10 матчів; кількість перемог безпосередньо перед матчем (стрік); рейтинг команди (унікальний); стабільність складу команди; середній показник програшів. В створеній ШНМ є три шари. Перший шар складається із 10 нейронів – рецепторів, або нейронів вхідних даних. Другий шар нейронів є внутрішній. Третій шар нейронів є вихідний, в ньому є лише 2 нейрони (рис. 3б).

6 ОГОВОРЕННЯ

Змагання з кіберспорту, у тому числі і міжнародні, проводяться по всьому світі. Найзначнішим з них є турнір orld Cyber Games, організований подібно до оімпійських ігор. Існує багато букмекерських ресурсів з прогнозуванням результатів матчів, але вони всі орієнтовані на прийом грошових ставок, їхні прогнози базуються на думці окремих букмекерів [1]. Інує багато порталів для перегляду подій зі світу кіберспорту, можливостями проведення ставок та публікування прогнозів на кіберспорт від експетів, наприклад:

1. http://cyber.sports.ru/;
2. https://r3.cybbet.com/;
3. http://bookmakerclub.com/articles_stavki_na_kibersport;
4. https://bookmaker-ratings.com.ua/tip/sport-types/tip-prognozy-na-kiber-sport/.

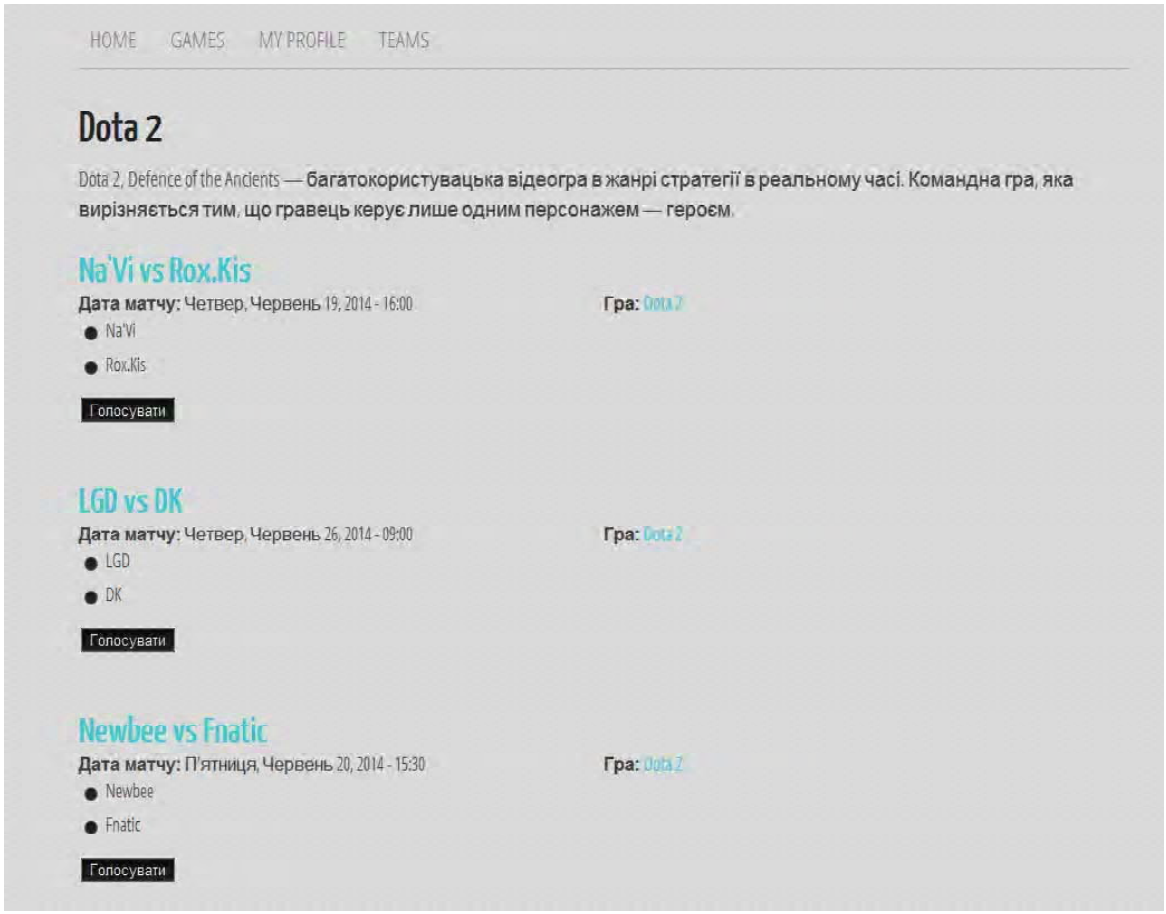
Але не існує як загального алгоритму проведення ефективних прогнозів в кіберспорті автоматично. Зазвичай прогнози роблять певне коло фахівців цієї сфери. Крім того, не існує загальної архітектури для систем прогнозування матчів у кіберспоті. Авторами розроблено та описано в роботі загальну архітектуру для систем прогнозування матчів у кіберспоті, в якій реалізовано розроблений авторами алгоритму проведення ефективних прогнозів матчів.

Спроектована ІС спрямована на задоволення потреб фанатів кіберспортивних ігор для відсліковування матчів. Кіберспорт розвивається дуже активно. Це викликає великий попит на кіберспортивний контент. Прогнозування результатів ігор є невід'ємною частиною цього контенту. Дана ІС орієнтована в на пересічних гравців і фанів кіберспорту. Подібних ІС майже немає. Одна з них egamingbets.com є подібною на розроблювану ІС, але вона орієнтована не на прогнозування результатів кіберсортивних матчів, а на виконання ставок на переможців цих ігор. Мінусом даної системи є також непрактичний інтерфейс, який важко зрозуміти новому користувачу, також відсутнє сортування ігор по даті, що є дуже зруч-

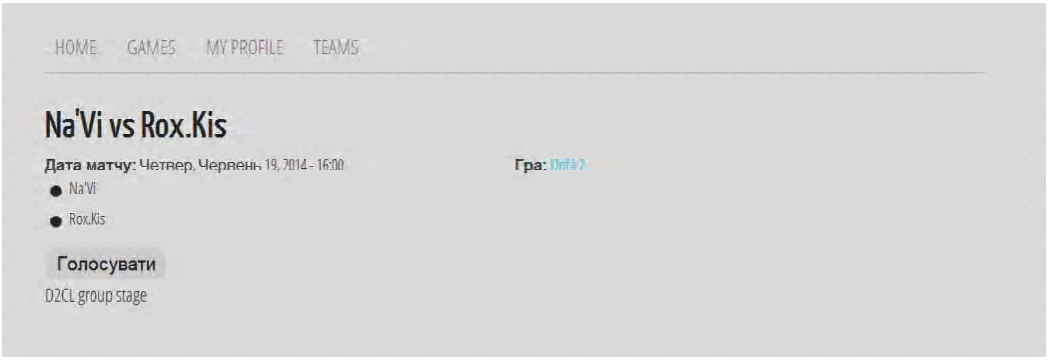
Рисунок 9 – Вікно прогнозування

Таблиця 1 – Урок для нейронної мережі

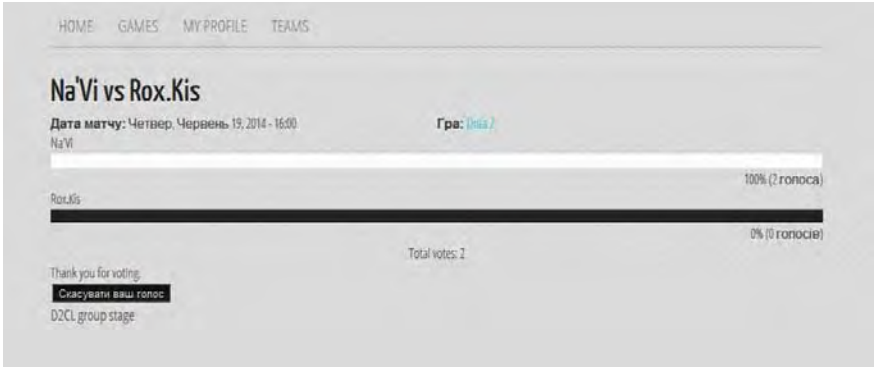
Команда 1					Команда 2					Переможець
В.з10	Стрік	Рейтинг	Склад	ККД	В.з10	Стрік	Рейтинг	Склад	ККД	
6	3	6554	3	68	5	5	6544	2	66	2
4	0	6234	1	63	6	3	6444	3	74	2
8	3	6592	2	87	7	6	6531	2	58	1
7	6	6587	3	68	6	4	6594	3	63	2
5	0	6498	2	74	6	2	6475	1	74	1
10	10	6612	3	59	6	4	6597	3	55	1
8	7	6537	2	66	6	6	6530	3	86	2
6	4	6498	1	71	5	5	6430	3	90	2



а



б



в

Рисунок 10 – Сторінка:
а – матчів; б – прогнозування; в – результатів прогнозування з гри

ною функцією. Значною перевагою даної системи також є те, що вона орієнтована на роботу не з однією грою а декількома такими як: Dota2; World of Tanks; Counter-Strike:GO. Список модулів CMS Drupal, які використовувались при розробці системи.

ВИСНОВКИ

В статті вирішено завдання розроблення загальної архітектури системи колективного прогнозування результатів ігор у кібер-спорті з використанням сучасної технології нейропрогнозування. Під час дослідження спроектовано, розроблено та апробувано функціонування такої системи.

Наукова новизна полягає у реалізації методу колективного прогнозування результатів ігор у кібер-спорті на основі розробленого алгоритму розрахунку прогнозів та навчання ШНМ. Реалізація такого методу підтримує незалежний від людського фактору процес прогнозування матчів в кібер-спорті.

Практична цінність розробленої ІС полягає у значному спрощенні пошуку прогнозів на кібер-спотривні матчі та дає можливість кожному бажаючому прийняти участь у прогнозуванні матчів. Система не є унікальною, проте вона може дати новий поштовх до вирішення проблеми прогнозування результатів не лише у кіберспорті, а і у спорті взагалі. Також перевагою даної системи є те, що вона безкоштовна, в той час, як за букмекерські прогнози зазвичай треба платити.

Перспективи подальших досліджень полягає у вдосконаленні алгоритму прогнозування кібер-матчів та методу навчання розробленої ШНМ на основі зібраних статистичних даних функціонування цієї ІС.

ПОДЯКИ

У статті розв'язана науково-практична задача прогнозування матчів у кіберспоті. Роботу виконано в рамках спільних наукових досліджень кафедри інформаційних систем та мереж Національного університету «Львівська політехніка» на тему «Дослідження, розроблення і впровадження інтелектуальних розподілених інформаційних технологій та систем на основі ресурсів баз даних, сховищ даних, просторів даних та знань з метою прискорення процесів формування сучасного інформаційного суспільства», а також Військово-дипломатичної академії імені Євгена Березняка. Наукові дослідження провадилися також в рамках ініціативної тематики досліджень кафедри ІСМ Національного університету «Львівська політехніка» на тему «Розроблення інтелектуальних розподілених систем на основі онтологічного підходу з метою інтеграції інформаційних ресурсів».

СПИСОК ЛІТЕРАТУРИ

1. Берко А. Ю. Системи електронної контент-комерції: монографія / А. Ю. Берко, В. А. Висоцька, В. В. Пасічник. – Львів : Видавництво Національного університету «Львівська політехніка», 2009. – 612 с.
2. Гаркуша Н. М. Моделі і методи прийняття рішень в аналізі та аудиті / Н. М. Гаркуша. – Х. : Знання, 2011. – 364 с.
3. Завгородня Т. Методи колективних експертних оцінок [Електронний ресурс] / Т. Завгородня. – Режим доступу : http://lubbook.net/book_251_glava_9_Tema_5._Metodi_kolektivnikhe.html.
4. Сазанов В. Г. Прогнозирование и планирование в условиях рынка / В. Г. Сазанов. – Владивосток : ТИДОТ ДВГУ, 2001. – 267 с.
5. Мамолов К. О. Методы экспертных оценок в планировании и прогнозировании / К. О. Мамолов. – Санкт-Петербург : ВНИИМП, 1999. – 289 с.
6. Уоссермен Ф. Нейрокомпютерная техника / Ф. Уоссерман. – М. : Мир, 1992. – 183 с.
7. Новаторський М. А. Штучні нейронні мережі : обчислення / М. А. Новаторський, Б. Б. Нестеренко. – К. : Ін-т математики НАН України, 2004. – 408 с.
8. Математична лінгвістика / [В. Висоцька, В. Пасічник, Ю. Щербина, Т. Шестакевич]. – Л. : «Новий Світ-2000», 2012. – 359 с.
9. Methods based on ontologies for information resources processing : monograph / [Vasyl Lytvyn, Victoria Vysotska, Lyubomyr Chyrun, Dmytro Dosyn]. – Saarbrücken : LAP. – 324 p.
10. Висоцька В. А. Методи і засоби опрацювання інформаційних ресурсів в системах електронної контент-комерції : автореферат дисертації на здобуття наукового ступеня кандидата технічних наук : 05.13.06 – інформаційні технології / Вікторія Анатоліївна Висоцька ; Національний університет «Львівська політехніка». – Львів, 2014. – 27 с.

Стаття надійшла до редакції 29.01.2017.
Після доробки 21.03.2017.

Коробчинский М. В.[1], Чирун Л. Б.[2], Высоцкая В. А.[3], Ныч Н. А.[4]

[1]Д-р. техн. наук, старший научный сотрудник Военно-дипломатической академии имени Евгения Березняка, Киев, Украина

[2]Ведущий специалист института компьютерных наук и информационных технологий Национального университета «Львовская политехника», Украина

[3]Канд. техн. наук, доцент, доцент кафедры «Информационные системы и сети» Национального университета «Львовская политехника», Украина

[4]Магистр кафедры «Информационные системы и сети» Национального университета «Львовская политехника», Украина

ОСОБЕННОСТИ ПРОГНОЗИРОВАНИЯ РЕЗУЛЬТАТОВ МАТЧЕЙ В КИБЕРСПОРТЕ

Актуальность. Сейчас актуальны разработки систем прогнозирования матчей в киберспоте, что связано с активным развитием киберспорта. В этой статье реализовано возможность прогнозировать матчи пользователей.

Цель. Целью выполнения работы является проектирование модели системы коллективного прогнозирования результатов игр в кибер-спорте с использованием современной технологии нейропрогнозирования. Задачей является разработка системы для совместного пользовательского прогнозирования результатов киберспортивных матчей и самостоятельной работы информации и выдачи собственного прогноза. К основным задачам относятся следующие: учет и анализ всех прошлых и будущих игр; учет и анализ характеристик / результатов всех команд; предоставление возможности пользователю персонально делать прогноз на каждый матч; определение шанса на выигрыш команды на основе данных за предыдущие матчи.

Метод. Проблема решена методом опроса экспертов путем проведения аналитических записок и с помощью искусственной нейронной сети. В созданной нейросети есть три слоя. Первый слой состоит из 10 нейронов-рецепторов, или нейронов входных данных. Второй слой нейронов является внутренним. Третий слой нейронов является выходным, в нем есть 2 нейроны. Входной

информацией для алгоритма является количество выигранных матчей из последних 10; количество выигранных матчей перед данной встречей (винстрик), рейтинг команды; стабильность состава (время неизменности состава команды), средний показатель проигрышей данной команды. Ответом является 1 или 2 (победа конкретной команды).

Результаты. Для достижения результата проведен анализ соответствующей литературы с информацией об основных видах коллективного прогнозирования. Разработано дерево цели, и проведен систематический анализ предметной области. Применен метод интервью с экспертами. Интернет-ресурсы реализованы CMS Drupal. Проанализированы основные методы коллективного прогнозирования. Проведенный систематический анализ объекта исследования и предмета, целей, построенных деревьев, определены проблемы и построено UML-диаграммы. Проанализировано применение метода интервью с экспертами. Реализовано сайт с CMS Drupal и языка программирования PHP.

Выводы. На основе разработанного алгоритма расчета прогнозов и обучение нейросети реализовано независимый от человеческого фактора процесс прогнозирования матчей в кибер-спорте. Наличие такой системы значительно упростит поиск прогнозов на кибер-спортивные матчи и даст возможность каждому желающему принять участие в их прогнозировании. Система дает новый толчок к решению проблемы прогнозирования результатов не только в кибер-спорте, а и в спорте вообще.

Ключевые слова: киберспорт, информационная система.

Korobchynskyi M. V.[1], Chyrun L. B.[2], Vysotska V. A.[3], Nych M. O.[4]

[1]Dr. Sc., Senior Research Fellow of Military-Diplomatic Academy named by Eugene Bereznyak, Kyiv, Ukraine

[2]Leading Specialist of Information Systems and Networks Department of Lviv Polytechnic National University, Lviv, Ukraine

[3]PhD, Associate Professor, Associate Professor of Information Systems and Networks Department of Lviv Polytechnic National University, Lviv, Ukraine

[4]Master of Information Systems and Networks Department of Lviv Polytechnic National University, Lviv, Ukraine

MATCHES PROGNOSTICATION FEATURES AND PERSPECTIVES IN CYBERSPORT

Context. The forecasting system fixing in cybersport is relevant at this point, it is connected to an active development cybersport. In this paper the ability to predict matches users is implemented.

Objective. The purpose of this work is the system model design of collective predicting outcomes of games results in cyber sport using modern technology of the neuro forecasting. The task is a system development for common user forecasting results of matches cybersport and independent information processing and issuing its own forecast. The main tasks include: all past and future games accounting and analysis; all teams performance / results accounting and analysis; user capabilities to make personal prediction for every match; determining the team winning odds based on previous matches.

Method. The problem by the survey of experts by means of analytical reports and using artificial neural network is solved. ANN is established in three layers. The first layer consists of 10 neurons-receptors, neurons of inputs data. The second layer of neurons is inside. The third layer is the source of neurons, it is only 2 neurons. The input data for the algorithm is the number of matches won the last 10; the number of won games before this meeting; team rating; stability of (time invariance of the team); the average of the losing team. The answer is 1 or 2 (the victory of a particular team).

Results. To achieve the cops of the relevant literature with information about the main types of collective prediction is reviewed. A tree objectives, and conducted a systematic analysis of the future system is developed. In this paper the method of interviews with experts and implemented it in their system is altered. Online resources implemented with CMS Drupal. The basic methods of collective prediction are described. A systematic analysis of the research object and subject is developed; objectives tree is developed and the problem and constructed UML-diagrams are studied. The method of interviews with experts and the use of this method in the paper are described. The implementation of a web site with CMS Drupal and programming language PHP is showed.

Conclusions. Based on the algorithm calculation example and training artificial neural network independent of the human factor in the process of predicting matches cyber sport is implemented. The presence of such system greatly simplifies the search for example cyber sports matches and give everyone the opportunity to take part in predicting matches. This system can give new impetus to the problem of predicting results not only in cyber-sport and sport in general.

Keywords: cybersport, prediction, information system.

REFERENCES

1. Berko A. YU. Vysotska V. A., Pasichnyk V. V. Systemy elektronnoyi kontent-komertsiyi: monohrafiya. L'viv, Vydavnytstvo Natsional'noho universytetu «L'vivs'ka politekhnika», 2009, 612 p.
2. Harkusha N. M. Modeli i metody pryynyattya rishen' v analizi ta audyti. KH, Znannya, 2011, 364 p.
3. Zavgorodnya T. Metodi kolektivnikh yekspertnikh otsinok [Electronic recourse]. Access mode : http://lubbook.net/book_251_ glava_9_Tema_5._Metodi_kolektivnikhe.html. The name on the title screen.
4. Sazonov V. G. Prognozirovaniye i planirovaniye v usloviyakh rynka. Vladivostok, TIDOT DVGU, 2001, 267 p.
5. Mamolov K. O. Metody ekspertnykh otsenok v planirovanii i prognozirovanii. Sankt-Peterburg, VNIIMP, 1999, 289 p.
6. Uossermen F. Neyrokompyuternaya tekhnika. Moscow, Mir, 1992, 183 p.
7. Novators'kyy M. A., Nesterenko B. B. Shtuchni neyronni merezhi: obchyslennya. K. In-t matematyky NAN Ukrayiny, 2004, 408 p.
8. Pasichnyk V. V., Shcherbyna YU. M., Vysotska V. A., Shestakevych T. V. Matematychna lingvistyka. [Knyha 1. Kvantytatyvna lingvistyka] : navch. posibnyk, *Seriya «Komp'yutyng»*. L'viv : «Novyy svit -2000», 2012, 359 p.
9. Vasyl Lytvyn, Victoria Vysotska, Lyubomyr Chyrun, Dmytro Dosyn Methods based on ontologies for information resources processing : monograph. Saarbr ücken, LAP, 324 p.
10. Vysotska V. A. Metody i zasoby opratsyuvannya informatsiynykh resursiv v systemakh elektronnoyi kontent-komertsiyi : avtoreferat dysertatsiyi na zdobuttya naukovoho stupenya kandydata tekhnichnykh nauk : 05.13.06 – informatsiyni, Natsional'nyy universytet «L'vivs'ka politekhnika». L'viv, 2014, 27 p.

ПРОГРЕСИВНІ ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

ПРОГРЕССИВНЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

PROGRESSIVE INFORMATION TECHNOLOGIES

УДК 519.161

Данильченко А. О.

Старший викладач кафедри комп'ютерної інженерії Житомирського державного технологічного університету, Житомир, Україна

РОЗПАРАЛЕЛЮВАННЯ МОДИФІКОВАНОГО МЕТОДУ ГІЛОК ТА МЕЖ ДЛЯ РОЗВ'ЯЗАННЯ ЗАДАЧІ ПРО ПАРОСПОЛУЧЕННЯ ЗІ ЗНИКАЮЧИМИ ДУГАМИ

Актуальність. Розглянуто задачу складання розкладу проходження процедур пацієнтами санаторію, яка зведена до розширеної задачі пошуку максимального паросполучення в дводольному графі. Для поставленої задачі про паросполучення зі зникаючим дугами було розроблено оптимальний алгоритм її рішення на базі методу гілок і меж. Алгоритм враховує обмеження сумісності процедур. Проведено розрахунковий експеримент в основі якого лежить доказ доцільності розпаралелювання оптимального алгоритму розв'язання задачі складання розкладу прийому лікувальних процедур пацієнтами для прикладного використання його в санаторних закладах України.

Мета роботи. Довести доцільність розпаралелювання оптимального алгоритму розв'язання задачі складання розкладу проходження процедур пацієнтами санаторію.

Метод. Сформульована математична модель задачі про паросполучення зі зникаючим дугами. Обрані обчислювальні платформи різної конфігурації, що мають різні обчислювальні потужності: різну кількість ядер процесора, різний обсяг пам'яті, і т.д. Написано авторське програмне забезпечення для проведення експерименту. Програма складається з двох модулів: серверний модуль, який контролює процес виконання розрахунків і клієнтський модуль, який виконується на відокремлених ПЕОМ з метою обчислення паралельних операцій. Проведено обчислювальний експеримент по распараллеливанию оптимального алгоритму розв'язання задачі про паросполучення зі зникаючим дугами. Експеримент проводився на базі санаторію «Деніші». Обчислювальний експеримент проведений на серії випадкових умов задачі, що генеруються програмою. Проведено аналіз отриманих результатів шляхом порівняння часу рішення задачі про паросполучення зі зникаючим дугами оптимальним алгоритмом на різних обчислювальних платформах.

Результати. Модифікований метод гілок та меж показує стабільність зменшення часу складання розкладу проходження процедур при збільшенні обчислювальних потужностей.

Висновки. Прогнозований найменший час складання розкладу, отримано на обчислювальній платформі з максимальною кількістю задіяних ПЕОМ. Прогнозований час складання розкладу при використанні алгоритму розпаралелювання модифікації методу гілок і меж прямо пропорційно залежить від кількості вершин дводольного графа (що дорівнює сумі кількості процедур і кількості пацієнтів), кількості призначених процедур і обмежень.

Ключові слова: паросполучення, дводольний граф, метод гілок і меж, метод повного перебору, розпаралелювання.

НОМЕНКЛАТУРА

ЗПІД – задача про призначення з умовами несумісності деяких пар робіт та їх виконавців;

$k_{ЛП}$ – кількість лікувальних процедур;

$k_{П}$ – кількість пацієнтів;

$k_{ПП}$ – кількість призначених процедур;

$C_{i,j}$ – обмеження прийому процедур;

G – дводольний орграф;

X – множина вершин, кожна з яких відповідає можливому проміжку прийому процедур;

Y – множина вершин відповідних множині всіх процедур;

ICS_ DENISH – програма автоматизованого формування розкладу прийому процедур пацієнтами санаторію;

$t_{ВР}^{ПР}$ – прогнозований час формування розкладу;

$t_{МП}$ – час складання розкладу методом повного перебору.

ВСТУП

Призначення процедур в сучасних санаторних та лікувальних закладах є складним процесом, що повинен враховувати досить велику кількість факторів, основними з яких є [1, 2]:

– перелік призначених лікарем процедур;

– час роботи процедурного кабінету;

DOI 10.15588/1607-3274-2017-3-12

– пропускна здатність процедурного кабінету (одну процедуру одночасно можуть приймати декілька пацієнтів);

– тривалість прийому процедури (для різних процедур тривалість прийому процедури різна);

– тривалість часу технічної перерви між прийомами процедур;

– сумісність процедур (пацієнт не може одночасно приймати декілька процедур, але крім цього, на розклад накладається додаткове обмеження – пацієнт не може приймати наступну процедури менш ніж через деякий час після прийняття попередньої, для кожної пари процедур значення часу сумісності може різнитися).

Такі задачі з успіхом розв'язуються математичним апаратом теорії розкладу – розділ прикладної математики, що вивчає моделі впорядкування та методи складання розкладів [3].

В статті [1] розглянуто задачу складання розкладу проходження процедур пацієнтами санаторію. Сформульована задача зведена до розширеної задачі пошуку максимального паросполучення у дводольному графі. Розроблено оптимальний алгоритм її розв'язання на базі відомого методу гілок та меж .

Із NP-повноти задачі «Про паросполучення зі зникаючими дугами», яка доведена в [1], слідує, що вона не піддається ефективним точним методам [4].

Відомо, що будь-яка NP-повна задача може бути розв'язана методом повного перебору. Але при цьому, в залежності від розмірності задачі, потрібні обчислювальні ресурси та час її розв'язання можуть бути неприпустимо великими з практичної точки зору.

Для оптимізації процесу повного перебору застосовують метод гілок та меж, який дозволяє зменшувати множину допустимих розв'язків за допомогою ефективного алгоритму пошуку, а також розпаралелювання обчислень.

Слід зазначити, що для методу віток та меж найбільшу складність мають саме процедура розгалуження та процедура знаходження оцінок верхніх і нижніх меж для оптимального значення на підмножині припустимих розв'язків.

Розпаралелювання обчислень не звужує кількість варіантів, що аналізуються, а лише скорочує потрібний на це час.

Таким чином, найбільш доцільною схемою зменшення обчислювальної складності знаходження точного розв'язання NP-повних задач залишається скорочення повного перебору.

Метою обчислювального експерименту є визначення доцільності розпаралелювання запропонованої модифікації методу гілок та меж для розв'язання задачі про паросполучення зі зникаючими дугами для прикладного використання його в санаторних закладах України.

1 ПОСТАНОВКА ЗАДАЧІ

Оскільки програмний продукт є орієнтований на кінцевого покупця загальна постановка задачі має наступний вигляд:

1. Потрібно розробити задачу, що забезпечує збереження таких даних:

– відомості про пацієнтів та збереження архіву даних;

– повна характеристика процедур та час сумісності процедур;

– інформація про лікарів їх спеціалізації та вагові коефіцієнти;

– вихідні та свята.

2. Задача має забезпечити виконання наступних функцій та розрахунків:

– додавання, редагування, знищення, сортування, друк та пошук даних по всім перерахованим пунктам;

– закріплення хворого за певним лікарем та розрахунок навантаження лікарів;

–призначення лікарем пацієнту певної кількості процедур;

– формування розкладу пацієнта з урахуванням завантаження процедурних кабінетів;

– система має забезпечити формування розкладу в двох режимах:

3. Формування розкладу пацієнта одразу після призначення (ручний режим);

4. Автоматичне формування розкладу для всіх пацієнтів.

– можливість редагування розкладу;

– розрахунок перегляд та друк загальних та вільних лімітів лікарів на призначення процедур;

– розрахунок, перегляд і друк резерву процедур та загальної кількості призначених процедур;

– розрахунок та друк завантаження процедурних кабінетів на певну дату.

В основі порівняльного обчислювального експерименту є доказ доцільності розпаралелювання алгоритму розв'язку задачі складання розкладу приймання лікувальних процедур пацієнтами для прикладного використання його в санаторних закладах України.

Для досягнення мети експерименту необхідно:

1. Оцінити часові витрати на виконання розрахунків на різних обчислювальних платформах, що мають різні обчислювальні потужності: різну кількість ядер мікропроцесора, різний обсяг пам'яті, тощо.

2. Порівняти результати та визначити ефективність та доцільність розпаралелювання процесу розв'язку. Критерієм ефективність є мінімізація часу виконання розрахунків щодо пошуку найбільшого паросполучення у дводольному графі зі зникаючими дугами.

2 ОГЛЯД ЛІТЕРАТУРИ

Метод гілок і меж [5] є загальним алгоритмічним методом вирішення різноманітних оптимізаційних задач. Він широко застосовується для таких NP-повних задач, як задача комівояжера та задача о ранці. В методі гілок і меж використовуються дві процедури: розгалуження та знаходження оцінок (меж).

Із NP-повноти задачі «Про паросполучення зі зникаючими дугами», яка доведена в [1], слідує, що вона не піддається ефективним точним методам [4].

Відомо, що будь-яка NP-повна задача може бути розв'язана методом повного перебору. Але при цьому, в залежності від розмірності задачі, потрібні обчислювальні ресурси та час її розв'язання можуть бути неприпустимо великими з практичної точки зору.

Для оптимізації процесу повного перебору застосовують метод віток та меж, який дозволяє зменшувати

множину допустимих розв'язків за допомогою ефективного алгоритму пошуку, а також розпаралелювання обчислень [6].

Слід зазначити, що для методу віток та меж найбільшу складність мають саме процедура розгалуження та процедура знаходження оцінок верхніх і нижніх меж для оптимального значення на підмножині припустимих розв'язків.

Розпаралелювання обчислень не звужує кількість варіантів, що аналізуються, а лише скорочує потрібний на це час.

У багатьох публікаціях запропоновано алгоритми рішення класичної задачі про паросполучення, які дозволили скоротити обчислювальну складність програмних реалізацій [7, 8].

3 МАТЕРІАЛИ І МЕТОДИ

Ряд задач про паросполучення в дводольних графах містить обмеження, що забороняє при виборі однієї дуги включати в рішення іншу дугу. Наприклад, якщо дуга *(x, y)* вже міститься в паросполучені, то йому не може належати дуга *(v, w)*. Такі дуги називаються несумісними. Вони включені в умови задачі про паросполучення зі зникаючим дугами (ЗПІД). Очевидно, ЗПІД - це задача про призначення з умовами несумісності деяких пар робіт та їх виконавців.

Розглянемо одну з прикладних версій ЗПІД – задача розподілу в часі оздоровчих процедур між пацієнтами санаторію. Санаторій надає список різних процедур. Пацієнт повинен пройти лікувальний курс з усіх або декількох процедур цього списку. Для кожної окремої процедури заданий графік її проведення в вигляді послідовності часових проміжків. Пацієнту призначається не більше однієї процедури зі списку. Потрібно знайти відповідність між множиною всіх процедур і множиною всіх часових проміжків:

Опишемо умови поставленої задачі в термінах дводольних графів. Нехай $G = (X, Y, E)$ – дводольний орграф, де X – множина вершин x_i, кожна з яких відповідає можливому проміжку прийому процедур, $i = 1, m$; Y – множина вершин y, $j = 1, n$ відповідних множині всіх процедур, що призначаються пацієнтам. Дуга $(x_i, y_j) \in E$, тоді і тільки тоді, коли процедуру y_j можна прийняти в проміжку часу x_i.

Очевидно, рішенням задачі є максимальне паросполучення в дводольному графі.

Доповнимо умови задачі наступним обмеженням: друга процедура призначається після першої не раніше ніж через годину. Тоді з паросполучення, що включає дугу (x_1, y_1), повинні автоматично зникнути дуга (x_5, y_2) не сумісна з дугою (x_1, y_1). У загальному випадку умова несумісності задана на деякій підмножині дуг двудольного орграфа. Потрібно знайти паросполучення, що включає максимальну кількість сумісних дуг.

В процесі покрокової побудови такого паросполучення деякі дуги будуть зникати, змінюючи структуру двудольного орграфа. Вже згадана ЗПІД – це задача про максимальне паросполучення в дводольному орграфе з заданою підмножиною дуг (x_i, y_j). Для кожної дуги (x_i, y_j) орграфа визначена підмножина дуг $C_{i,j}$, які не включаються до допустимого рішення, якщо в нього включена дуга (x_i, y_j). Дуги множини $C_{i,j}$ зникають з графа G при включенні дуги (x_i, y_j) і знову стають видимими при її виключенні. Відношення несумісності дуги (x_i, y_j) з підмножиною дуг $C_{i,j}$ позначимо $(x_i, y_j) \rightarrow C_{i,j} = \{(x_{i1}, y_{j1}), (x_{i2}, y_{j2}), \ldots, (x_{ik}, y_{jk})\}$.

Для розв'язку сформульованої задачі було розроблено оптимальний алгоритм на базі метода гілок та меж [2].

Для подальшого проведення експерименту було розроблено програму NetRemoting, що призначена для автоматизованого формування розкладу прийому процедур пацієнтами санаторію, який складається за допомогою оптимального алгоритму рішення задачі ЗПІД. Крім того, програма забезпечує: розрахунок навантаження лікарів, розрахунок, перегляд та друк загальних і вільних лімітів лікарів, розрахунок, перегляд і друк резерву процедур та загальної кількості призначених процедур, розрахунок завантаження процедурних кабінетів.

Програма автоматизованого формування розкладу прийому процедур пацієнтами санаторію (ICS_ DENISH) розрахована на виконання на ПЕОМ типу IBM PC і працює під управлінням операційної системи Windows 8 або вище. Операційна система, що рекомендується, – Windows 10.

Перед проведенням експерименту виконані наступні інструкції

1. Встановлено програмне забезпечення (netframework «v2.0.50727»).

2. Налаштовано ПЕОМ, на якій виконується серверний модуль:

2.1. Відключено антивірус.

2.2. Відключено Брандмауер.

2.3. Додано до виключень порт 8008, по якому йде обмін даними.

2.4. Визначено IP адресу сервера.

3. Налаштовано клієнтський програмний модуль:

3.1. В файл Client.exe.config вписано IP адресу сервера «tcp://192.168.2.101:8008/PllCompOfBB/ControlProcess.rem».

3.2. Відключено антивірус.

3.3. Відключено Брандмауер.

4 ЕКСПЕРИМЕНТИ

Експеримент проводився з використанням авторського проблемно-орієнтованого інструментарію – програми ICS_ DENISH. Програма складається з двох модулів: серверний модуль, що контролює процес виконання розрахунків та клієнтський модуль, який виконується на відділених ПЕОМ з метою обчислення паралельних операцій. Програма вирішує задачу складання розкладу приймання процедур пацієнтами санаторію модифікованим методом гілок та меж. Обчислювальний експеримент проведено на серії випадкових умов задачі, що генеруються програмою. Вхідні параметри наведені у табл. 2 були випадковими (реєстрованими, але некерованими). Вихідним параметром для проведення обчислювального експерименту є $t_{BP}^{ПР}$ – прогнозований час формування розкладу за таких умовах. Кількість випробувань та вхідні параметри відповідають (табл. 1–2).

Вихідним параметром для проведення порівняльного обчислювального експерименту є час $t_{ВР}$, що витрачається на виконання розрахунків щодо складання розкладу приймання лікувальних процедур пацієнтами санаторію за допомогою програми ICS_DENISH.

Узагальнені інструкції використання програмного забезпечення при проведенні обчислювального експерименту:

1. Запустити на сервері програмне забезпечення, обрати потрібну нам кількість клієнтів і чекати їх підключення. Основне вікно серверного модуля представлено на рис. 1.

2. Після підключення всіх клієнтів сервер переходить в режим генерації вихідних даних, вибираємо потрібну нам кількість процедур, людей, призначених процедур і натискаємо кнопку, сервер генерує дані. Вікно генерації вихідних даних представлено на рис. 2.

Сервер переходить до режиму готовності до обчислень. Вікно готовності серверного модуля до обчислень наведено на рис. 3.

3. Отримати результат. Екран програми з результатом обчислення наведено на рис. 4.

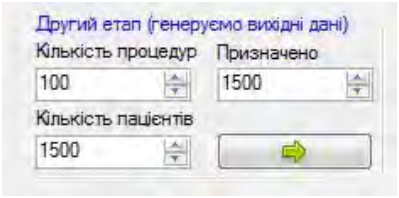


Рисунок 2 – Вікно генерації вихідних даних

Рисунок 4 – Вікно з результатом обчислень $t_{BP}^{ПР}$

Для проведення обчислювального експерименту були використані комп'ютери з характеристиками згідно табл. 1, з яких було сформовано три обчислювальні платформи, що наведені на рис. 5.

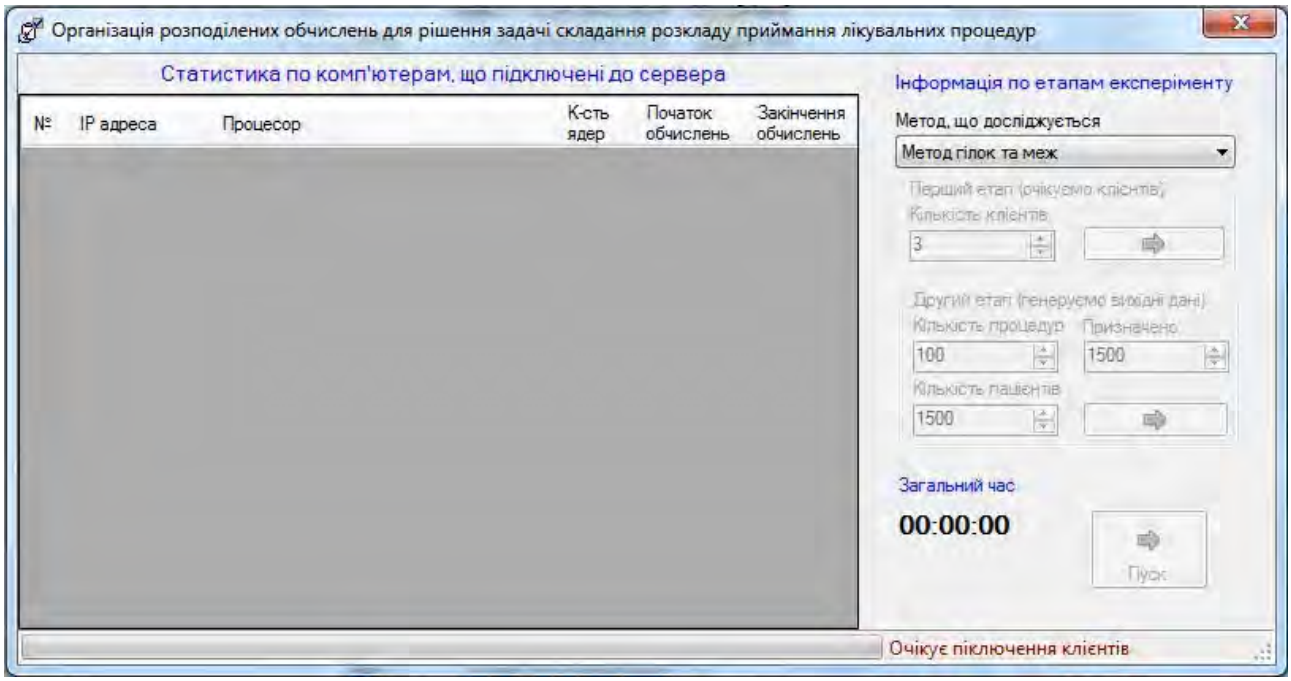


Рисунок 1 – Основне вікно серверного модуля

Таблиця 1 – Характеристики комп'ютерів, що були використані для проведення порівняльного обчислювального експерименту

Номер обчислювальної платформи	Назва мікропроцесора	Кіль. ядер	Тактова частота	Обсяг оперативної пам'яті ПЕОМ
1	Pentium 4	1	2,42 ГГц	512 МБ
2	Intel Celeron E3300	2	2,50 ГГц	2 ГБ
3	Intel Core i5-3570K	4	3,4 ГГц	4 ГБ

Таблиця 2 – Перелік вхідних параметрів обчислювального експерименту

Номер параметру	Позначка	Назва параметру	Тип параметру
1	$k_{ЛП}$	Кількість лікувальних процедур	Детермінований
2	$k_{П}$	Кількість пацієнтів	Випадковий
3	$k_{ПП}$	Кількість призначених процедур	Випадковий
4	$C_{i,j}$	Обмеження прийому процедур	Детермінований

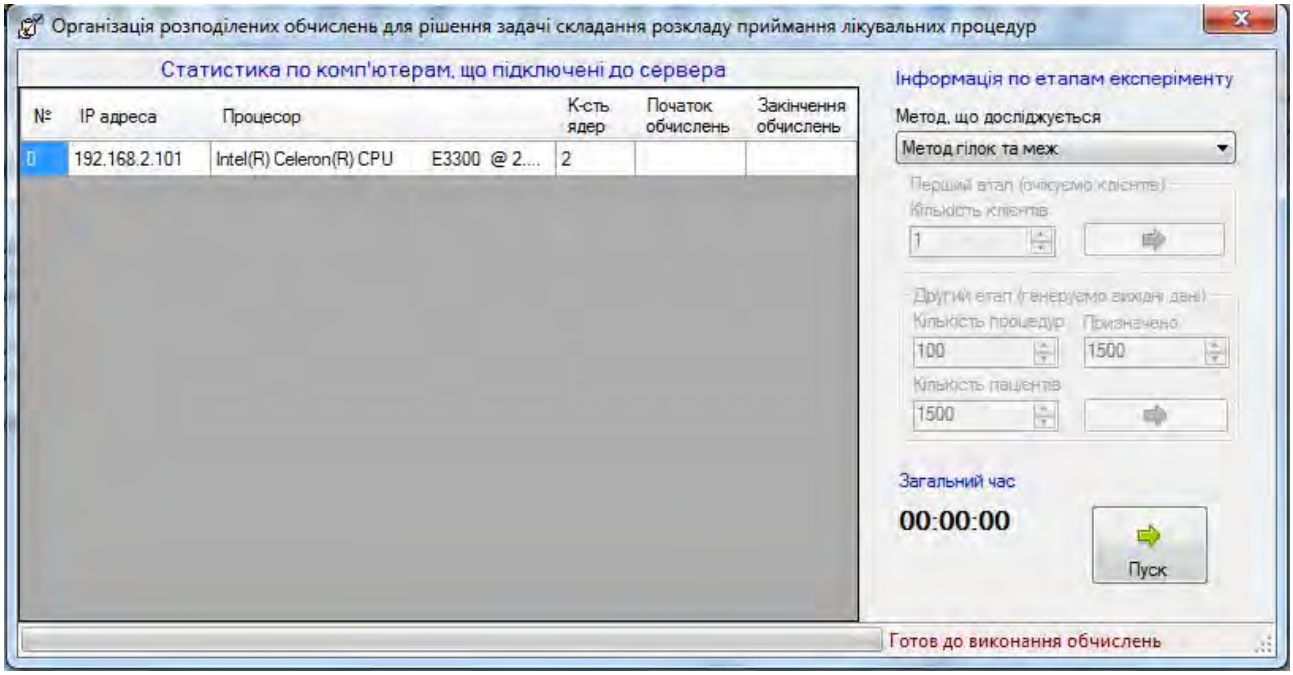
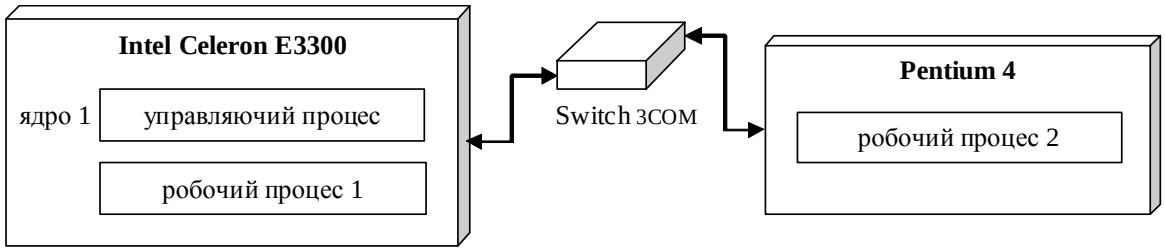
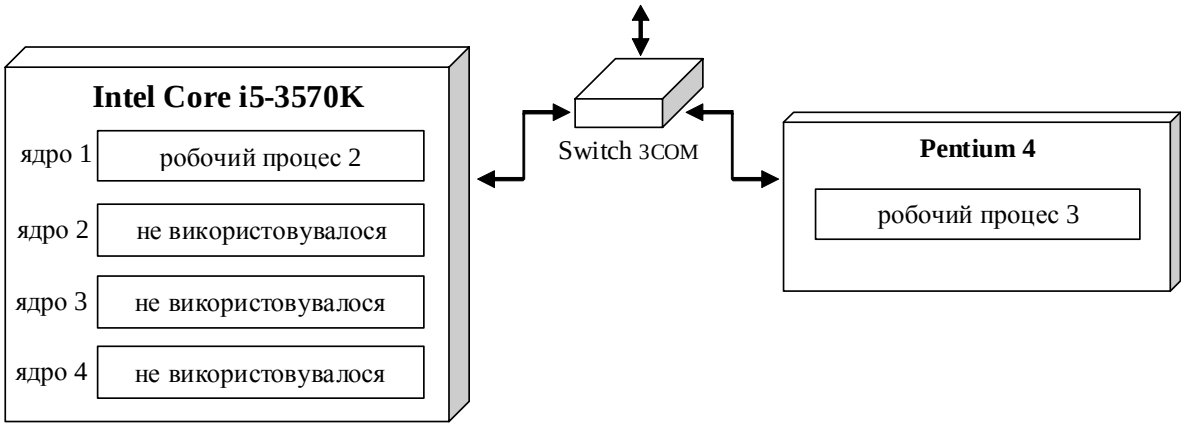


Рисунок 3 – Вікно готовності серверного модуля до обчислень

а



б



в

Рисунок 5 – Структура процесів, що були задіяні на різних обчислювальних платформах:
а – перша обчислювальна платформа; б – друга обчислювальна платформа; в – третя обчислювальна платформа

Таблиця 6 – Результати окремих випробувань обчислювального експерименту

Номер випробування	Вхідні параметри			Прогнозований час формування розкладу		
	$k_{ЛП}$	$k_{П}$	$K_{ПП}$	Перша платформа	Друга платформа	Третя платформа
1	74	1584	2974	71,2	37,4	23,8
2	74	1312	3217	91,5	48,8	31,2
3	80	1723	3343	112,4	59,2	38,7
4	80	1784	3578	115,8	60,1	40,4
5	85	1837	3724	137	71,7	47
…	…	…	…	…	…	…
86	129	8719	51622	412,2	209,3	138,1

5 РЕЗУЛЬТАТИ

У кожному випробуванні наведеному у табл. 6 наведено вимір часу на виконання розрахунків модифікацією методу гілок та меж для різних обчислювальних платформ, що наведені на рис. 5.

6 ОБГОВОРЕННЯ

Результати дослідження дозволяють зробити наступні висновки:

1. Прогнозований час складання розкладу при використанні алгоритму розпаралелювання модифікації методу гілок та меж прямо пропорційно залежить від кількості вершин дводольного графу (яка дорівнює сумі кількості процедур та кількості пацієнтів), кількості призначених процедур та обмежень.

2. Прогнозовано найменший час складання розкладу, який отримано на обчислювальній платформі з максимальною кількістю задіяних ПЕОМ.

3. Отримано максимальне зменшення часу в 19,1 разів в порівнянні 86-го випробування для третьої платформи з 86 випробуванням для першої платформи для методу повного перебору ($t_{МП}$).

4. Таким чином, доведено доцільність розпаралелювання запропонованої модифікації методу гілок та меж для розв'язання задачі про паросполучення зі зникаючими дугами для прикладного використання його в санаторних закладах України.

ВИСНОВКИ

1. Було проаналізована задача складання розкладу прийому процедур пацієнтами санаторію. Поставлена задача складання розкладу прийому лікувальних процедур сформульована в термінах теорії графів і з урахуванням заданих обмежень описана дводольним графом. Показано, що рішення задачі зводиться до знаходження максимального паросполучення в цьому графі.

2. Проведена модифікація задачі про паросполучення зі зникаючим дугами для складання розкладу при заданих обмеженнях.

3. Розроблено оптимальний алгоритм вирішення задачі про паросполучення, суть якого полягає в знаходженні всіх максимальних паросочетание максимальної потужності з подальшою їх перевіркою на сумісність із заданими обмеженнями.

4. Із застосуванням розробленого програмного продукту проведено порівняльний обчислювальний експеримент, який дозволив оцінити часові характеристики оптимального алгоритму розв'язання задачі про паросполучення зі зникаючим дугами на різних розрахункових платформах і порівняти їх. За результатами розрахункового моделювання, можна зробити висновки про доцільність розпаралелювання рішення задачі про паро сполучення із зникаючими дугами оптимальним алгоритмом. Модифікований метод гілок та меж показує стабільність зменшення часу складання розкладу приймання процедур пацієнтами при збільшенні обчислювальних потужностей ПЕОМ, Так, наприклад, при $k_{ЛП}$= 85, $k_{П}$ = 1837, $k_{ПП}$= 3724 час вирішення задачі на третій платформі зменшується у 3 рази порівняно з першою обчислювальною платформою.

Практична цінність досліджень полягає в можливості їх використання при розробці та застосуванні систем календарного планування і оперативного управління в лікувальному процесі. Дослідження також застосовні при розробці систем управління гнучкими автоматизованими системами для підприємств з дискретним характером виробництва.

Автори планують надалі додати ваги кожної дузі і вирішити задачу про паросполучення з «зникаючими» дугами в модифікованому вигляді.

Перспективним напрямком подальших досліджень є модифікація відомих методів (генетичного, мурашиного і т.д.) Для вирішення задачі складання розкладу прийому процедур пацієнтами санаторію.

СПИСОК ЛІТЕРАТУРИ

1. Данильченко А. О. Розв'язання одного класу задач складання розкладів генетичними алгоритмами на кластерних системах / О. М. Данильченко, А. О. Данильченко, С. А. Ібрагім // Вісник ЖІТІ. – 2004. – № 4. – С.130–135.
2. Данильченко А. О. Задача про паросполучення зі «зникаючими» дугами / А. О. Данильченко, А. В. Панішев, А. М. Данильченко // Збірник наукових праць «Моделювання та інформаційні технології». – 2012. – № 63. – С.75–81.
3. Лупин С. А. http://sevntu.com.ua/cgi-bin/irbis64r_72/cgiirbis_64.exe?Z21ID=&I21DBN=JOURN_PRINT&P21DBN=JOURN&S21STN=1&S21REF=&S21FMT=fullw_print&C21COM=S&S21CNR=&S21P01=0&S21P02=0&S21P03=M=&S21STR= Метод решения задач составления расписания, ориентированный на кластерные вычислительные системы / С. А. Лупин, Т. В. Милехина // Известия ВУЗов. Сер. Электроника. – 2007. – № 6. – С. 63–69
4. Пападимитриу Х. Комбинаторная оптимизация. Алгоритмы и сложность / Х. Пападимитриу, К. Стаглиц. – Москва : Мир, 1985. – 512 с.
5. Жолобов Д. А. Введение в математическое программирование: учебное пособие / Д. А. Жолобов. – Москва: МИФИ, 2008. – 376 с.
6. Агеев А. А. Приближенный алгоритм решения метрической задачи о двух коммивояжерах с оценкой точности / А. А. Агеев, А. В. Пяткин // Дискретный анализ и исследование операций. Серия 1 : Сибирское отделение Российской академии наук. Институт математики им. С. Л. Соболева СО РАН. – 2009. – Том 16, № 4. – С. 3–20.

7. Li Wenxia A DNA Algorithm for the Maximal Matching Problem / Li Wenxia, E. Patrikeev, Xiao Dongmei // Automatics and robot. – 2015. – № 10. – С. 106–112.
8. Sonkin D. Adaptive algorithm of distributing orders for taxi service / D. Sonkin // The Tomsk Polytechnic University. – 2009. – № 5. – С. 65–69.

Стаття надійшла до редакції 10.04.2017.
Після доробки 27.06.2017.

Данильченко А. А.
Старший преподаватель кафедры компьютерной инженерии Житомирского государственного технологического университета, Житомир, Украина

РАСПАРАЛЛЕЛИВАНИЕ МОДИФИКАЦИИ МЕТОДА ВЕТВЕЙ И ГРАНИЦ ДЛЯ РЕШЕНИЯ ЗАДАЧИ О ПАРОСОЧЕТАНИИ С ИСЧЕЗАЮЩИМИ ДУГАМИ

Актуальность. Рассмотрена задача составления расписания прохождения процедур пациентами санатория, которая сведена к расширенной задаче поиска максимального паросочетания в двудольном графе. Для поставленной задачи о паросочетании с исчезающими дугами разработан оптимальный алгоритм ее решения на базе метода ветвей и границ. Алгоритм учитывает ограничения совместимости процедур. Проведен расчетный эксперимент в основе которого лежит доказательство целесообразности распараллеливания оптимального алгоритма решения задачи составления расписания приема лечебных процедур пациентами для прикладного использования его в санаторных заведениях Украины.

Цель работы. Доказать целесообразность распараллеливания оптимального алгоритма решения задачи составления расписания прохождения процедур пациентами санатория.

Метод. Сформулирована математическая модель задачи о паросочетании с исчезающими дугами. Выбраны вычислительные платформы разной конфигурации имеющие различные вычислительные мощности: разное количество ядер процессора, разный объем памяти, и т.д. Написано авторское программное обеспечение для проведения эксперимента. Программа состоит из двух модулей: серверный модуль, контролирующий процесс выполнения расчетов и клиентский модуль, который выполняется на отделенных ПЭВМ с целью вычисления параллельных операций. Проведен вычислительный эксперимент по распараллеливанию оптимального алгоритма решения задачи о паросочетании с исчезающими дугами. Эксперимент проводился на базе санатория «Дениши». Вычислительный эксперимент проведен на серии случайных условий задачи, генерируемых программой. Проведен анализ полученных результатов путем сравнения времени решения задачи о паросочетании с исчезающими дугами оптимальным алгоритмом на разных вычислительных платформах.

Результаты. Модифицированный метод ветвей и границ показывает стабильность уменьшения времени составления расписания прохождения процедур при увеличении вычислительных мощностей.

Выводы. Прогнозируемое наименьшее время составления расписания, получено на вычислительной платформе с максимальным количеством задействованных ПЭВМ. Прогнозируемое время составления расписания при использовании алгоритма распараллеливания модификации метода ветвей и границ прямо пропорционально зависит от количества вершин двудольного графа (которое равно сумме количества процедур и количества пациентов), количества назначенных процедур и ограничений.

Ключевые слова: паросочетание, двудольный граф, метод ветвей и границ, метод полного перебора, распараллеливания.

Danylchenko A.
Senior lecturer in Computer Engineering Zhytomyr State Technological University, Zhitomir, Ukraine

PARALLELING MODIFIED METHOD OF BRANCH AND BOUND TO SOLVE PROBLEM OF MATCHING CURVES FROM ENDANGERED OR THREATENED

Context. The problem of scheduling the passage of procedures of sanatorium patients, which is reduced to the problem of finding an extended maximum matching in a bipartite graph. For the task of matchings with disappearing arcs developed an optimal algorithm of its solution based on branch and bound method. The algorithm takes into account the limits of compatibility procedures. Spend the current experiment based on the evidence of the feasibility of algorithm parallelization for solving the problem of optimal scheduling patients receiving therapeutic treatments applied to its use in the health institutions of Ukraine.

Objective. To prove the feasibility of the algorithm parallelization optimal solution of our problem.

Method. A mathematical model of the problem of matchings with disappearing arcs. Selected computing platforms of different configurations with a variety of computing power: a different number of processor cores, different amounts of memory, etc. Written copyright software for the experiment. The program consists of two modules: a server module, which controls the process of performing calculations and client module that runs on the PC are separated for the purpose of calculating the parallel operations. The experiment was conducted on the basis of sanatorium “Denyshi”. Computational experiments for optimal algorithm parallelization for solving the problem of matchings with disappearing arcs. Computer experiment carried out on a series of random conditions of the problem generated by the program. The analysis of the results by comparing the time solving the problem of matchings with disappearing arcs optimal algorithm on different computing platforms.

Results. The modified method of branches and borders shows the stability of reducing the time of scheduling transmission procedures with increasing computing power.

Conclusions. Estimated minimum time scheduling, received at the computer platform with the maximum number of PCs involved. Estimated time scheduling algorithm parallelization by using modifications of the branch and bound directly proportional to the number of vertices of a bipartite graph (which is equal to the sum of the number of procedures and the number of patients), the number of assigned procedures and restrictions.

Keywords: matching, bipartite graphs, branch and bound method, the method of exhaustive search, parallelization.

REFERENCES

1. Danil’chenko A. O., Danil’chenko A. O., Íbragím S. A. Rozv’yazannya odnogo klasu zadach skladannya rozkladív genetichnimi algoritmami na klasternikh sistemakh, *Vísnik ZHÍTÍ*, 2004, No. 4, pp. 130–135.
2. Danil’chenko A. O., Panísheva V. V., Danil’chenko A. M. Zadacha pro parospoluchennya zí «znikayuchimi» dugami, *Zbírnik naukovikh prats’ «Modelyuvannya ta ínformatsíyní tekhnologíí»*, 2012, No. 63, pp. 75–81.
3. Lupin S.. The method for solving scheduling problems, focused on cluster computing systems, Proceedings of the universities. Ser. Electronics: scientific-technical, 2007, 6, pp. 63–69.
4. Papadimitriu KH., Staglits K. Kombinatornaya optimizatsiya. Algoritmy i slozhnost’. Moscow, Mir, 1985, 512 p.
5. Zholobov D. A. Vvedeniye v matematicheskoye programmirovaniye: uchebnoye posobiye. Moscow, MIFI, 2008, 376 p.
6. Ageev A. Approximate algorithm for solving the problem of metric peripatetic salesman with an estimate of the accuracy.Discrete Analysis and Operations Research. Series 1: Siberian Branch of the Russian Academy of Sciences. Institute of Mathematics. Siberian Branch of the Russian Academy of Sciences. 2009, 4, pp. 3–20.
7. Li Wenxia Patrikeev E., Dongmei Xiao A DNA Algorithm for the Maximal Matching Problem, Automatics and robot, 2015, No. 10, pp. 106–112.
8. Sonkin D. Adaptive algorithm of distributing orders for taxi service, *The Tomsk Polytechnic University*, 2009, No. 5, pp. 65–69.

УДК 004.056

Имамвердиев Я. Н.[1], Сухостат Л. В.[2]

[1]Канд. техн. наук, доцент, зав. отделом, Институт информационных технологий Национальной Академии Наук Азербайджана, Баку, Азербайджан
[2]Канд. техн. наук, старший научный сотрудник, Институт информационных технологий Национальной Академии Наук Азербайджана, Баку, Азербайджан

ОБНАРУЖЕНИЕ АНОМАЛИЙ В СЕТЕВОМ ТРАФИКЕ НА ОСНОВЕ ИНФОРМАТИВНЫХ ПРИЗНАКОВ

Актуальность. Решена актуальная задача оценки информативности признаков данных большой размерности. Объектом исследования являлся сетевой трафик.

Цель работы – анализ данных сетевого трафика на предмет информативности для выявления аномалий в сетевом трафике с целью сокращения пространства признаков.

Метод. Предложен подход для оценки информативности признаков данных большой размерности, обеспечивающий повышение точности выявления аномалий в сетевом трафике и существенно увеличивающий скорость работы алгоритмов классификации. Проанализированы особенности алгоритмов случайного леса и Firefly. В работе для отбора признаков предложен подход на основе интеграции данных алгоритмов. Признаки сортируются в порядке убывания оценки их важности, наименее информативные не рассматриваются. В качестве классификаторов были рассмотрены деревья решений, наивный Байес, Байесовский классификатор, аддитивная логистическая регрессия и метод к-ближайших соседей. Результаты классификации были оценены с использованием пяти метрик: вероятности истинно-положительных и ложно-положительных результатов, F-меры, мер точности и полноты.

Результаты. Эксперименты были проведены в среде Matlab 2016a, где был реализован предложенный алгоритм на наборе данных NSL-KDD. Наилучшие результаты классификации для отобранных признаков были получены методом к-ближайших соседей.

Выводы. Проведенные эксперименты подтвердили работоспособность предложенного подхода, что позволяет рекомендовать его для применения на практике при оценке информативности с целью сокращения пространства признаков и повышения скорости работы алгоритмов классификации. Кроме того, в целях дальнейшего изучения эффективности обнаружения аномалий в сетевом трафике, будет использован набор реальных данных.

Ключевые слова: сетевые атаки, информативность признаков, случайный лес, алгоритм Firefly, NSL-KDD.

НОМЕНКЛАТУРА

RF – Random Forest;
KDD – Knowledge Discovery and Data Mining;
OOB – Out-of-Bag Error;
DoS – Denial of Service Attack;
U2R – Users to Root Attack;
R2L – Remote to Local Attack;
Weka – Waikato Environment for Knowledge Analysis;
Probe – Probing Attack;
TCP – Transmission Control Protocol;
UDP – User Datagram Protocol;
ICMP – Internet Control Message Protocol;
TPR – True Positive Rate;
FPR – False Positive Rate;
AUC – Area under ROC Curve;
BayesNet – Байесовский классификатор;
J48 – деревья решений;
LogitBoost – аддитивная логистическая регрессия;
IBk – метод к-ближайших соседей;
x_{ij} – точка из набора данных;
$\xi(x_{ij})$ – номер класса, которому соответствует значение x_{ij};
c – количество распознаваемых классов;
Θ – область допустимых значений;
I – информативность признака;
P – популяция светлячков;
β – привлекательность светлячка;
γ – коэффициент поглощения света;
r – расстояние;
α – параметр рандомизации;
T_i – ансамбль деревьев решений;
$\hat{\omega}_i(x)$ – класс новых наблюдений;
Q_t – распределение Гаусса случайных чисел.

ВВЕДЕНИЕ

В последнее время с развитием сетевых технологий угрозы безопасности значительно возросли [1, 2]. Таким образом, повышение уровня сетевой безопасности является одним из актуальных вопросов для исследователей [3].

При анализе данных сетевого трафика проблема размерности стоит остро. Размерность имеющихся данных, характеризующаяся различным числом признаков, достигает большого числа показателей. В силу этого необходимо снизить размерность признакового пространства и выделить из них наиболее важные.

Отбор признаков помогает улучшить производительность классификации и ее способности к обобщению. Другим мотивом для отбора признаков является то, что меньшее количество признаков приводит к более интерпретируемым классификаторам, что важно во многих областях (например, биомедицине).

Кроме того, измерение некоторых переменных признаков может быть довольно дорогостоящим с точки зрения денег, требований к хранению и передаче данных или времени на обучение. Данные с меньшей размерностью также могут быть более легко визуализированы. Таким образом, отбор признаков является важной задачей во многих системах классификации образов.

DOI 10.15588/1607-3274-2017-3-13

Во многих исследованиях был сделан вывод о том, что различные алгоритмы отбора признаков имеют различное поведение на различных наборах данных, и, следовательно, опасно использовать только один алгоритм [4].

Методы машинного обучения широко применяются для отбора признаков с целью анализа сетевого трафика на предмет наличия атак. Теоретически алгоритмы машинного обучения могут получить высокую производительность, т.е. могут минимизировать уровень ложных тревог и максимизировать точность обнаружения. Однако обычно требуется бесконечное число обучающих образцов. На практике это условие невозможно в силу ограничения вычислительной мощности и требования ответа в режиме реального времени.

Существует множество алгоритмов работающих на основе имитации поведения природных агентов, таких как рыбы, птицы, насекомые и т.д. Среди них алгоритм Firefly (алгоритм «светлячков») является одним из тех, который может приводить к эффективным решениям большого числа задач [5]. Целью данного исследования является разработка нового подхода для отбора признаков путем интеграции алгоритмов случайного леса (random forest, RF) [6, 7] и Firefly.

1 ПОСТАНОВКА ЗАДАЧИ

Для оценки информативности в работе рассматриваются алгоритмы случайного леса и Firefly, на основе которых отбираются наиболее важные признаки [8].

Обозначим через Θ область допустимых значений. Строки матрицы $X \in R^{m \times n}$ при этом представляют элементы обучающей выборки, $\xi(x_{ij})$ – номер класса, которому соответствует значение x_{ij} j-го признака на i-ом элементе выборки, а c – количество распознаваемых классов. Далее производится оценка информативности $I(x_{ij})$ $(i = 1,..., m)$ j-го признака с областью определения Θ алгоритмом случайного леса. Признаки сортируются в порядке убывания оценки их важности, наименее информативные не рассматриваются.

Далее на основе алгоритма Firefly необходимо сгенерировать популяцию светлячков P, где каждый светлячок соответствует отобранным признакам. При этом необходимо определить изменчивость интенсивности света (variation of light intensity) и формулировку привлекательности (attractiveness formulation). Привлекательность светлячка пропорциональна интенсивности света, которая меняется с расстоянием r и задается в виде,

$$\beta = \beta_0 e^{-\gamma r^2}, \qquad (1)$$

где β_0 превращается в привлекательность при $r = 0$. Движение светлячка k привлекает другого более яркого светлячка l и определяется как

$$y_k^{t+1} = y_k^t + \beta_0 e^{-\gamma r_{kl}^2}(y_l^t - y_k^t) + \alpha t Q_k^t. \qquad (2)$$

Требуется оценить информативность признаков сетевого трафика для повышения скорости работы систем обнаружения вторжений, сохраняя при этом достаточно хорошие результаты.

2 ЛИТЕРАТУРНЫЙ ОБЗОР

При классификации набор данных обычно включает большое количество признаков, которые могут быть релевантными, нерелевантными или избыточными. Избыточные и нерелевантные признаки не пригодны для классификации, и они могут даже снизить эффективность классификатора в отношении большого пространства поиска, которое также известно как «проклятие размерности» [9].

Преимущества отбора признаков включают в себя сокращение вычислительных затрат, экономию дискового пространства, упрощение процедур выбора модели для точного прогнозирования и интерпретации комплексных зависимостей между переменными [10]. Отобранные признаки не только оптимизируют точность классификации, но также уменьшают количество необходимых данных для достижения оптимального уровня производительности процесса обучения [11, 12].

Методы отбора признаков обычно включают в себя стратегию поиска, меру оценки, критерий остановки и валидацию результатов.

Среди двух подходов, используемых для отбора признаков, а именно метода фильтров (filter approach) и метода обертки (wrapper approach), первый работает лучше при анализе данных высокой размерности [11].

Генетический алгоритм является одним из недавних разработок для отбора признаков [13]. В настоящее время он является очень эффективным в научно-технической оптимизации.

Классификация на основе протоколов была предложена с использованием генетического алгоритма с логистической регрессией и применена к набору данных KDD'99 в работах [14, 15].

Гибридный метод для отбора признаков при обнаружении сетевых вторжений представлен в работе [16]. В этой статье, речь идет о новом алгоритме, который сочетает в себе прирост информации и генетический алгоритм.

В [17] представлен современный подход для отбора признаков на основе алгоритма Firefly.

3 МАТЕРИАЛЫ И МЕТОДЫ

В данном разделе приводится описание алгоритмов случайного леса и Firefly.

Случайный лес был предложен Л. Брейманом в статье [6]. Он строится на основе ансамбля деревьев решений, каждый элемент которого получается при помощи бутсрепа (bootstrap) [18, 19]. Называется ансамблем по той причине, что при создании одного дерева используются не все признаки пространства, а только случайно выбранные.

Алгоритм случайного леса заключается в следующем:

Пусть обучающий набор состоит из m образцов, размерность пространства признаков при этом равна n. Строится необходимое число деревьев. С помощью голосования проводится классификация. Объект классификации будет отнесен каждым деревом к одному из классов, и класс, за который проголосовало большее количество деревьев, побеждает.

Для каждого дерева выбирается подвыборка из числа наблюдений и подвыборка из числа переменных [20]. На этой подвыборке обучается дерево.

Далее получается ансамбль деревьев решений $\{T_i\}_{i=1}^{s}$, где s – количество деревьев в ансамбле ($i = 1,2,...,s$).

При предсказании новых наблюдений получается класс $\hat{\omega}_i(x) \in \{\omega_1, \omega_2, ..., \omega_c\}$, предсказанный T_i, т.е. $T_i(x) = \hat{\omega}_i(x)$; где $\hat{\omega}_{rf}^{s}(x)$ – класс, наиболее часто встречающийся в множестве $\{\hat{\omega}_i(x)\}_{i=1}^{s}$ [21].

Для этой задачи можно использовать любые классификаторы, но деревья обладают способностью быстро обучаться. На основе метрики out-of-bag error (OOB) определяется ошибка [22–24].

Преимущества случайных лесов включают:

– значительное повышение точности;

– высокая вычислительная эффективность;

– переподгонка в некоторых случаях решаема (когда количестве признаков больше числа наблюдений в обучающей выборке);

– метод прост в применении.

Их недостатками являются отсутствие наглядного представления процесса принятия решения, а, следовательно, и сложность в интерпретации результатов.

Брейман предложил меры информативности признаков, что позволило строить матрицу близости наблюдений для компенсации перечисленных выше недостатков. Одной из важных задач статистического анализа является нахождение наиболее информативных признаков. А меры информативности дают такую возможность.

Алгоритм Firefly был предложен Xin She Yang и основан на поведении светлячков [5]. Основной алгоритм Firefly предполагает, что существует P светлячков y_k ($k = 1,..., p$), первоначально произвольно размещенных в пространстве. Интенсивность света I каждого светлячка определяется целевой функцией $f(x)$. В простейшей форме, интенсивность света $I(r)$ изменяется в зависимости от расстояния r монотонно и экспоненциально, как это показано в (3):

$$I = I_0 e^{-\gamma r}, \quad (3)$$

где I_0 – исходная интенсивность света и γ – коэффициент поглощения света. Если $I_i > I_j$, $j \neq i$, то менее яркий светлячок j будет двигаться в направлении более яркого светлячка i.

Привлекательность изменяется в зависимости от расстояния $r_{ij} = d(y_i, y_j)$:

$$r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}. \quad (4)$$

Движение одного светлячка к другому, более привлекательному светлячку определяется как (2). Если яркость j больше, чем i, то передвигаем i к j. Таким образом, сходимость светлячка определяется его движением.

В (2) второе слагаемое обусловлено привлекательностью. В третьем члене α – параметр рандомизации, а Q_t представляет собой распределение Гаусса случайных чисел. Если $\beta_0 = 0$, то движение становится произвольным. Если $\gamma = 0$, то задача сводится к оптимизации роя частиц.

4 ЭКСПЕРИМЕНТЫ

Для проведения экспериментов была рассмотрена база данных сигнатур NSL-KDD [25], построенная на основе базы KDD-99 по инициативе американской Ассоциации перспективных оборонных научных исследований DARPA [26]. Она охватывает широкий спектр различных вторжений, смоделированных в среде, имитирующей сеть Военно-воздушных сил США.

Рассмотренная база NSL-KDD имеет следующие преимущества [27]:

– нет избыточных записей в обучающем наборе, так что классификатор не покажет какой-либо предвзятый результат;

– нет дубликата записей в тестовом наборе. Он содержит некоторые атаки, которые не присутствуют в обучающем наборе;

– количество выбранных записей из каждой группы уровней сложности обратно пропорционально доле записей в исходном наборе данных NSL-KDD.

Обучающий набор данных состоит из 21 различных атак, а тестовый – из 37. Известные виды атак содержатся в обучающем наборе, в то время как новые атаки – это дополнительные атаки в тестовом наборе данных (они отсутствуют в обучающем наборе). Кроме того, количество записей в обучающем наборе составляет 125973 образцов, а в тестовом – 22544. Это преимущество делает его доступным для проведения экспериментов на полных данных без необходимости случайным образом выбирать небольшую часть.

Все атаки в NSL-KDD поделены на четыре группы [28]: DoS (Denial of Service Attack), U2R (Users to Root Attack), R2L (Remote to Local Attack) и Probe (Probing Attack).

Каждая запись имеет 42 атрибута, описывающих различные признаки (табл. 1). Протоколы, которые рассматриваются в NSL-KDD, включают TCP, UDP (User Datagram Protocol) и ICMP (Internet Control Message Protocol).

Метки присваиваются каждой записи либо в качестве типа «атаки», либо как «нормальное» состояние [29].

Для сравнения производительности и эффективности методов обнаружения вторжений в сети используются следующие метрики [30]:

а) Наиболее распространенными метриками для сравнения систем обнаружения вторжений являются вероятности истинно-положительных (True Positive Rate, TPR) и ложно-положительных результатов (False Positive Rate, FPR). FPR является вероятностью получения оповещения, даже если система ведет себя нормально. С другой стороны, вероятность ложно-отрицательных результатов (False Negative Rate, FNR) является вероятностью не дающей сигнала тревоги, даже если поведение системы является вредоносным. Уравнения (5) и (6) представляют FPR и FNR:

$$FPR = \frac{number\ of\ \ false\ positives}{number\ of\ \ false\ positives + number\ of\ true\ negatives}, \quad (5)$$

Таблица 1 – Список признаков для каждой записи базы данных NSL-KDD

№	Название признака	№	Название признака	№	Название признака
1	duration	15	su_attempted	29	same_srv_rate
2	protocol_type	16	num_root	30	diff_srv_rate
3	service	17	num_file_creations	31	srv_diff_hast_rate
4	flag	18	num_shells	32	dst_host_count
5	scr_bytes	19	num_access_files	33	dst_host_srv_count
6	dst_bytes	20	num_outbound_cmds	34	dst_host_same_srv_rate
7	land	21	is_host_login	35	dst_host_diff_srv_rate
8	wrong_fragments	22	is_quest_login	36	dst_host_same_src_port_rate
9	urgent	23	count	37	dst_host_srv_diff_host_rate
10	hot	24	srv_count	38	dst_host_serror_rate
11	num_failed_logins	25	serror_rate	39	dst_host_srv_serror_rate
12	logged_in	26	srv_serror_rate	40	dst_host_rerror_rate
13	num_compromised	27	rerror_rate	41	dst_host_srv_rerror_rate
14	root_shell	28	srv_rerror_rate	42	class

$$FNR = \frac{number\ of\ \ false\ negatives}{number\ of\ \ false\ negatives + number\ of\ true\ positives}. \quad (6)$$

Следовательно, вероятности TPR и истинно-отрицательных результатов (True Negative Rate, TNR) могут быть определены как:

$$TPR = 1 - FNR \ \text{ и } \ TNR = 1 - FPR. \quad (7)$$

По сути, существует компромисс между скоростью ложных срабатываний и частотой ложных отрицательных значений. Если политика обнаружения вторжений становится очень чувствительной, риск FPR будет выше. Таким образом, баланс следует рассматривать между этими двумя рисками (FPR и FNR) в конфигурации системы обнаружения вторжений.

б) Выражение (8) представляет собой меру полноты (recall), которая определяется как доля нормального поведения:

$$\text{recall} = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + number\ of\ \ false\ negatives}. \quad (8)$$

Тем не менее, мера полноты недостаточно содержательна, так как она может быть получена тривиальным образом путем классификации всех типов поведения, как вредоносных.

в) Существует еще одна метрика, называемая мерой точности (precision), которая решает эту проблему:

$$precision = \frac{number\ of\ true\ negatives}{number\ of\ true\ negatives + number\ of\ \ false\ positives}. \quad (9)$$

При классификации всего трафика как нормального, мера точности достигается полностью.

г) F-мера является показателем, который сочетает в себе меры точности и полноты:

$$F - measure = \frac{2 \times recall \times precision}{recall + precision}. \quad (10)$$

д) ROC-кривая (Receiver Operator Characteristic – рабочая характеристика (приемника)) сравнивает частоту TPR с FPR [30]. Одно важное ограничение этой метрики состоит в том, что она вычисляет общую производительность системы обнаружения вторжений на всех исходных данных. Чем выше значение площади под ROC-кривой (area under ROC curve, AUC), тем лучше производительность метода.

Разрабатываемый подход был проанализирован с использованием следующих классификаторов:

– наивного Байесовского классификатора (NaiveBayes);

– деревьей решений (J48);

– аддитивной логистической регрессии (Additive Logistic Regression – LogitBoost);

– Байесовского классификатора (BayesNet);

– метода к-ближайших соседей (IBk).

В ходе тестирования были использованы реализации данных алгоритмов в программной системе Weka 3.8.0.

5 РЕЗУЛЬТАТЫ

Эксперименты были проведены в ОС Windows® 10–64 с процессором Core i7 (2,5 ГГц), 8.0 Гб ОЗУ. Оценка информативности признаков проводилась в среде Matlab 2016a на наборе данных NSL-KDD. Параметры алгоритмов Firefly и случайного леса, использованные в эксперименте, приведены в табл. 2.

В результате ошибка алгоритма случайного леса составила 0,08% при количестве деревьев равном 30 и значении OOB равном 0,03%, что показывает хорошую работу подхода.

Пять различных алгоритмов классификации (деревья решений, наивный Байес, Байесовский классификатор, аддитивная логистическая регрессия и метод к-ближайших соседей) сравниваются в программной среде Weka 3.8.0 при отобранных информативных признаках (Таблица 3) и 41 признаке базы данных NSL-KDD.

В таблицах 4 и 5 приводятся результаты работы алгоритмов классификации сетевых атак на основе 41 признака и отобранных информативных признаков из имеющегося набора данных соответственно.

Наилучшие результаты были отмечены жирным шрифтом (табл. 4–5). В качестве метрик были рассмотрены TPR, FPR, Precision, Recall, F-мера и AUC.

Таблица 2 – Параметры алгоритмов оценки информативности признаков

Число светлячков	7
Параметр рандомизации (α)	0,2
Привлекательность	2
Коэффициент поглощения света (γ)	1
Число деревьев	30

Из табл. 4–5 можно сделать заключение, что наилучшие результаты были получены методом IBk. Согласно метрике FPR наименьший процент ошибки классификации для 41 признака был достигнут методом NaiveBayes, а для отобранных 25 признаков – методом IBk.

Сравнение производительности алгоритмов (Таблица 4) показало, что метод BayesNet превосходит остальные по метрике AUC (92,5%). Анализируя полученные данные по метрике F-мера уменьшение размерности вектора признаков согласно информативности привело к улучшению работы методов J48, NaiveBayes, LogitBoost и IBk.

Сравнение значений AUC для рассмотренных классификаторов более наглядно демонстрируется на рис. 1 (красным цветом обозначены результаты для 41 признака, а синим – 25 признаков на основе предложенного подхода).

Таблица 3 – Результаты оценки информативности признаков

Подход	Отобранные признаки
Алгоритм Firefly	22,26,30,5,37,14,7,13,27,21,10,18,24,23,11,12,31,1,20,36
Случайный лес	5,2,23,24,36,10,8,4,34,40,31,32,35,22,6,27,1,33,37,16,14,11,29,13,28
Предлагаемый подход	22,26,5,2,23,24,36,37,14,13,27,21,10,18,11,12,31,1,20,8,29,28,40,35,6

Таблица 4 – Сравнение производительности алгоритмов классификации для 41 признака

Метод	TPR (%)	FPR (%)	Precision (%)	Recall (%)	F-мера (%)	AUC (%)
J48	75,8	13,2	76,7	75,8	**74,0**	81,8
NaiveBayes	70,2	**11,7**	75,8	70,2	70,7	86,5
BayesNet	71,5	19,2	78,6	71,5	67,0	**92,5**
LogitBoost	74,5	15,8	78,0	74,5	73,3	90,6
IBk	**76,8**	16,3	**81,2**	**76,8**	72,6	82,3

Таблица 5 – Сравнение производительности алгоритмов классификации для отобранных признаков

Метод	TPR (%)	FPR (%)	Precision (%)	Recall (%)	F-мера (%)	AUC (%)
J48	76,5	14,7	80,9	76,5	74,3	82,1
NaiveBayes	59,5	7,9	73,4	59,5	64,9	84,7
BayesNet	73,9	16,9	80,2	73,9	69,8	93,6
LogitBoost	78,8	11,5	81,8	78,8	78,7	93,5
IBk	**99,6**	**0,2**	**99,6**	**99,6**	**99,6**	**100**

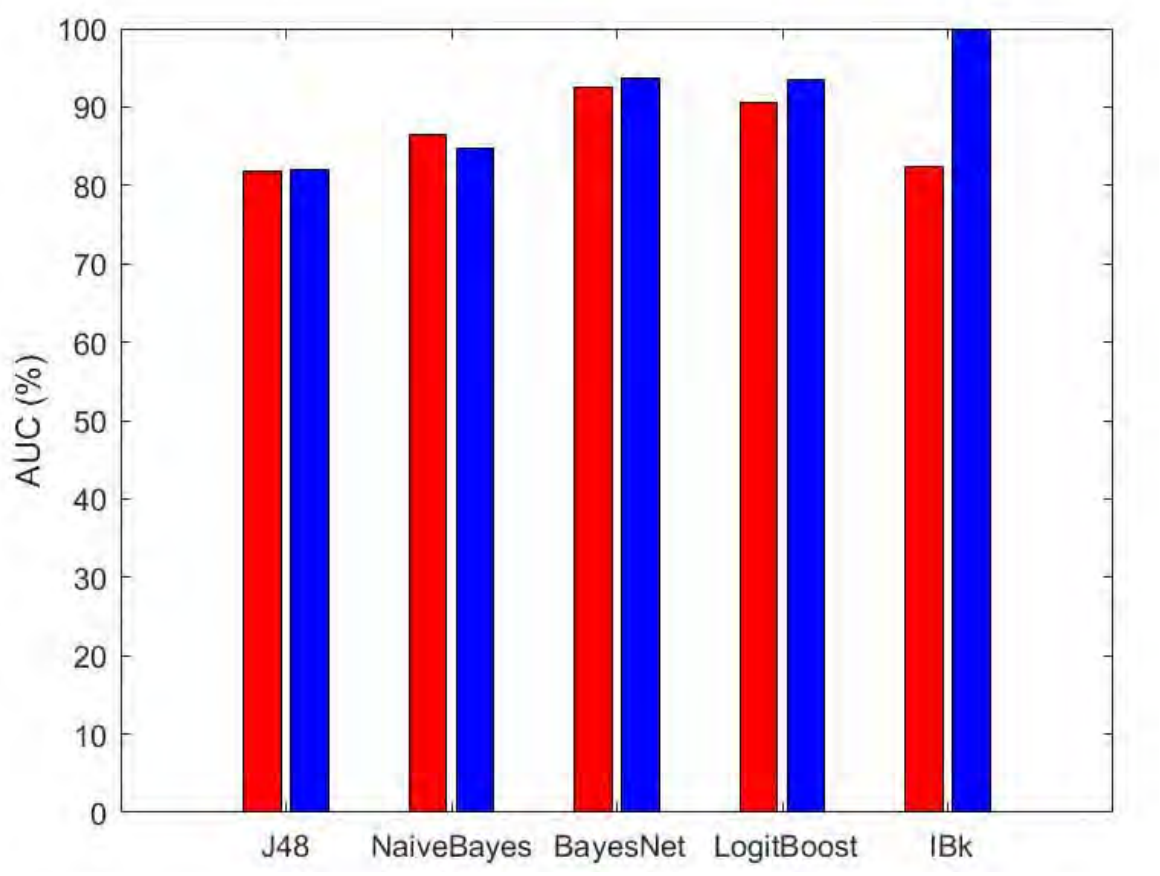


Рисунок 1 – Сравнение производительности методов классификации для базы данных NSL-KDD

6 ОБСУЖДЕНИЕ

Предложенный подход оценки информативности признаков данных большой размерности обеспечивает повышение точности выявления аномалий в сетевом трафике. Соответственно, сокращение пространства признаков существенно увеличивает скорость работы алгоритмов классификации.

Тестирование подхода проводилось на основе метрик TPR, FPR, Precision, Recall, F-мера и AUC. Наилучшие результаты были получены для метода к-ближайших соседей, однако он требует больших временных затрат. В силу этого предпочтение можно было бы отдать методам BayesNet или LogitBoost.

ВЫВОДЫ

Цель текущего исследования состояла в анализе данных высокой размерности на предмет информативности для выявления аномалий в сетевом трафике. В работе для выявления информативных признаков, используемых для обнаружения атак в сетевом трафике, были рассмотрены алгоритмы случайного леса и Firefly. В качестве классификаторов были рассмотрены деревья решений, наивный Байес, Байесовский классификатор, аддитивная логистическая регрессия и метод к-ближайших соседей.

Таким образом, экспериментальные результаты показывают, что предлагаемый подход достигает перспективной производительности при обнаружении сетевых атак на основе информативных признаков.

Хотя предложенный алгоритм отбора информативных признаков имеет обнадеживающую производительность, она может быть дополнительно повышена за счет оптимизации стратегии поиска. Кроме того, в целях дальнейшего изучения эффективности обнаружения аномалий в сетевом трафике, будет использован набор реальных данных.

СПИСОК ЛИТЕРАТУРЫ

1. Dua S. Data mining and machine learning in cybersecurity / S. Dua, X. Du. – Boca Raton, FL: CRC Press, 2011. – 256 p. DOI: 10.1201/b10867
2. Saxe J. Why security data science matters and how its different: pitfalls and promises of data science based breach detection and threat intelligence [Electronic resource]. – 2015. – Access mode: https://www.blackhat.com/us-15/speakers/Joshua-Saxe.html
3. Gates C. Challenging the anomaly detection paradigm: a provocative discussion / C. Gates, C.Taylor // Proceedings of the Workshop on New Security Paradigms. – 2007. – P. 21–29. DOI: 10.1145/505202.505211
4. Molina L. C. Feature selection algorithms: a survey and experimental evaluation / L. C. Molina, L. Belanche, A. Nebot // Proceedings of IEEE International Conference on Data Mining. – 2002. – P. 306–313. DOI: 10.1109/ICDM.2002.1183917
5. Yang X.-S. Firefly algorithms for multimodal optimization / X.-S. Yang // Stochastic Algorithms: Foundations and Applications. – 2009. – Vol. 5792. – P. 169–178. DOI: 10.1007/978-3-642-04944-6_14
6. Breiman, L. Random forests / L. Breiman // Machine Learning. – 2001. – № 1. – P. 5–32. DOI: 10.1023/A:1010933404324
7. Random forests – Classification manual [Electronic resource]. – 2017. Access mode: http://www.math.usu.edu/adele/Forests/
8. Strobl, C. Danger: High power! – exploring the statistical properties of a test for random forest variable importance / C. Strobl, A. Zeileis // Proceedings in Computational Statistics. – 2008. – P. 59–66.
9. Xue B. Particle swarm optimization for feature selection in classification: Novel initialization and updating mechanisms / B. Xue, M. Zhang, W. N. Browne // Applied Soft Computing. – 2014. – Vol. 18. – P. 261–276. DOI: 10.1109/TSMCB.2012.2227469
10. Feng D. Supervised feature subset selection with ordinal optimization / D. Feng, F. Chen, W. Xu // Knowledge-Based Systems. – 2014. – Vol. 56. – P. 123–140. DOI: 10.1016/j.knosys.2013.11.004
11. Bouaguel W. A fusion approach based on wrapper and filter feature selection methods using majority vote and feature weighting / W. Bouaguel, G. B. Mufti, M. Limam // Proceedings of the International Conference on Computer Applications Technology. – 2013. – P. 1–6. DOI: 10.1109/ICCAT.2013.6522003
12. Wang G. An improved boosting based on feature selection for corporate bankruptcy prediction / G. Wang, J. Ma, S. Yang // Expert Systems with Applications. – 2014. – Vol. 41, № 5. – P. 2353–2361. DOI: 10.1016/j.eswa.2013.09.033
13. Srinivasa K. G. Application of Genetic Algorithms for Detecting Anomaly in Network Intrusion Detection Systems / K. G. Srinivasa // Advances in Computer Science and Information Technology. Networks and Communications. – 2012. – Vol. 84. – P. 582–591. DOI: 10.1007/978-3-642-27299-8_61
14. Yu K. M. Protocol-based classification for intrusion detection / K. M. Yu, M. F. Wu, W. T. Wong // Applied Computer and Applied Computational Science. – 2008. – Vol. 3, № 3. – P. 135–141.
15. Akbar S. Intrusion detection system methodologies based on data analysis / S. Akbar, R. K. Nageswara, J. A. Chandulal // International Journal of Computer Applications. – 2010. – Vol. 5, № 2. – P. 10–20. DOI: 10.5120/892-1266
16. Sethuramalingam S. Hybrid feature selection for network intrusion detection / S. Sethuramalingam, E. R. Naganathan // International Journal of Computer Science and Engineering. – 2011. – Vol. 3, № 5. – P. 1773–1780. DOI: 10.4225/75/57a84d4fbefbb
17. Banati H. Fire Fly based feature selection approach / H. Banati, M. Bajaj // IJCSI International Journal of Computer Science Issues. – 2011. – Vol. 8, № 4. – P. 473–80.
18. Hothorn T. Unbiased recursive partitioning: a conditional inference framework / T. Hothorn, K. Hornik, A. Zeileis // Journal of Computational and Graphical Statistics. – 2006. – Vol. 15, № 3. – P. 651–674. DOI: 10.1198/106186006X133933
19. Breiman L. Stacked Regressions / L. Breiman // Machine Learning. – 1996. – Vol. 24. – P. 49–64. DOI: 10.1007/BF00117832
20. Strobl C. Conditional variable importance for random forests / C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, A. Zeileis // BMC Bioinformatics. – 2008. – Vol. 9, № 1. – P. 25. DOI: 10.1186/1471-2105-9-307
21. Siroky D. Navigating Random Forests and related advances in algorithmic modeling / D. Siroky // Statistics Surveys. – 2009. – Vol. 3. – P. 147–163. DOI: 10.1214/07-SS033
22. Archer K. J. Empirical characterization of random forest variable importance measures / K. J. Archer, R. V. Kimes // Computational Statistics & Data Analysis. – 2008. – № 4. – P. 2249–2260. DOI: 10.1016/j.csda.2007.08.015
23. Strobl C. Bias in random forest variable importance measures: illustrations, sources and a solution / C. Strobl, A.-L. Boulesteix, A. Zeileis, T. Hothorn // BMC Bioinformatics. – 2007. – Vol. 8, № 1. – P. 1471–2105. DOI: 10.1186/1471-2105-8-25
24. Liaw A. Classiыcation and Regression by randomForest / A. Liaw, M. Wiener // R News. – 2002. – Vol. 2, № 3. – P. 18–22.
25. Aggarwal P. Analysis of KDD dataset attributes-class wise for intrusion detection / P. Aggarwal, S. K. Sharma // Procedia Computer Science. – 2015. – Vol. 57. – P. 842–851. DOI: 10.1016/j.procs.2015.07.490
26. McHugh J. Testing intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations

as performed by lincoln laboratory / J. McHugh // ACM Transactions on Information and System Security. – 2000. – Vol. 3, № 4. – P. 262–294. DOI: 10.1145/382912.382923
27. Tavallaee M. A detailed analysis of the KDD CUP 99 Data Set / M. Tavallaee, E. Bagheri, W. Lu, A. Ghorbani // Proceedings of the second IEEE Symposium on Computational Intelligence for Security and Defense Applications. – 2009. – P. – 53–58. DOI: 10.1109/CISDA.2009.5356528
28. NSL-KDD data set for network-based intrusion detection systems [Electronic resource]. – 2017. – Access mode: http://nsl.cs.unb.ca/NSL-KDD/
29. Davis J. J. Data preprocessing for anomaly based network intrusion detection: A review / J. J. Davis, A. J. Clark // Computers & Security. – 2011. – Vol. 30, № 6–7. – P. 353–375. DOI: 10.1016/j.cose.2011.05.008
30. Holz T. 13 security measurements and metrics for networks / T. Holz // Dependability Metrics. – 2008. – P. 157–165. DOI: 10.1007/978-3-540-68947-8_13

Статья поступила в редакцию 07.04.2017.
После доработки 26.06.2017.

Имамвердиев Я. Н.[1], Сухостат Л. В.[2]
[1]Канд. техн. наук, доцент, зав. відділом, Інститут інформаційних технологій Національної Академії Наук Азербайджану, Баку, Азербайджан
[2]Канд. техн. наук, старший науковий співробітник, Інститут інформаційних технологій Національної Академії Наук Азербайджану, Баку, Азербайджан

ВИЯВЛЕННЯ АНОМАЛІЙ У МЕРЕЖЕВОМУ ТРАФІКУ НА ОСНОВІ ІНФОРМАТИВНИХ ОЗНАК

Актуальність. Вирішено актуальне завдання оцінки інформативності ознак даних великої розмірності. Об'єктом дослідження був мережевий трафік.

Мета роботи – аналіз даних мережевого трафіку на предмет інформативності для виявлення аномалій в мережевому трафіку з метою скорочення простору ознак.

Метод. Запропоновано підхід для оцінки інформативності ознак даних великої розмірності, що забезпечує підвищення точності виявлення аномалій в мережевому трафіку і істотно збільшує швидкість роботи алгоритмів класифікації. Проаналізовано особливості алгоритмів випадкового лісу і Firefly. В роботі для відбору ознак запропонований підхід на основі інтеграції даних алгоритмів. Ознаки сортуються в порядку убування оцінки їх важливості, найменш інформативні не розглядаються. Як класифікаторів були розглянуті дерева рішень, наївний Байес, Байєсівський класифікатор, аддитивная логістична регресія і метод до найближчих сусідів. Результати класифікації були оцінені з використанням п'яти метрик: ймовірності істинно-позитивних і хибно-позитивних результатів, F-заходи, заходів точності і повноти.

Результати. Експерименти були проведені в середовищі Matlab 2016a, де був реалізований запропонований алгоритм на наборі даних NSL-KDD. Найкращі результати класифікації для відібраних ознак були отримані методом к-найближчих сусідів.

Висновки. Проведені експерименти підтвердили працездатність запропонованого підходу, що дозволяє рекомендувати його для застосування на практиці при оцінці інформативності з метою скорочення простору ознак і підвищення швидкості роботи алгоритмів класифікації. Крім того, з метою подальшого вивчення ефективності виявлення аномалій в мережевому трафіку, буде використаний набір реальних даних.

Ключові слова: мережеві атаки, інформативність ознак, випадковий ліс, алгоритм Firefly, NSL-KDD.

Imamverdiyev Y. N.[1], Sukhostat L. V.[2]
[1]PhD, Associate Professor, Head of Department, Institute of Information Technology of Azerbaijan National Academy of Sciences, Baku, Azerbaijan
[2]PhD, Senior Researcher, Institute of Information Technology of Azerbaijan National Academy of Sciences, Baku, Azerbaijan

NETWORK TRAFFIC ANOMALIES DETECTION BASED ON INFORMATIVE FEATURES

Context. The urgent task for feature informativeness evaluation of a large amount of data has been solved. The object of the study was a network traffic.

Objective is to analyze the data informativeness for network traffic anomalies detection in order to reduce the feature space.

Method. The approach for feature informativeness evaluation of a large amount of data is proposed to increase the accuracy of the anomaly detection in network traffic. It also substantially increases the computation speed of the classification algorithms. The characteristics of a random forest and Firefly algorithms are considered. In the paper, an algorithm for feature selection based on the integration of these algorithms is proposed. Features are sorted in descending order according to their importance, the least informative ones are not considered. The decision trees, naive Bayes, Bayesian classifier, additive logistic regression and k-nearest neighbors method are considered as classifiers. The quality of the classification results is estimated using six evaluation metrics: true positive rate, false positive rate, precision, recall, F-measure and AUC.

Results. The experiments have been performed in the Matlab environment (2016a) on the NSL-KDD data set, using the proposed algorithm. The best classification results for the selected features have been obtained using k-nearest neighbors method.

Conclusions. The conducted experiments have confirmed the efficiency of the proposed approach and allow recommending it for practical use in feature informativeness evaluation in order to reduce the feature space and increase the computation speed of the classification algorithms. In addition, in order to further study the effectiveness of anomaly detection in network traffic, a real data set will be used.

Keywords: network attacks, feature informativeness, random forest, Firefly algorithm, NSL-KDD.

REFERENCES

1. Dua S., Du X. Data mining and machine learning in cybersecurity. Boca Raton, FL, CRC Press, 2011, 256 p. DOI: 10.1201/b10867
2. Saxe J. Why security data science matters and how its different: pitfalls and promises of data science based breach detection and threat intelligence [Electronic resource], 2015, Access mode: https://www.blackhat.com/us-15/speakers/Joshua-Saxe.html
3. Gates C., Taylor C. Challenging the anomaly detection paradigm: a provocative discussion, *Proceedings of the Workshop on New Security Paradigms*, 2007, pp. 21–29. DOI: 10.1145/505202.505211
4. Molina L. C., Belanche L., Nebot A. Feature selection algorithms: a survey and experimental evaluation, *Proceedings of IEEE International Conference on Data Mining*, 2002, pp. 306–313. DOI: 10.1109/ICDM.2002.1183917
5. Yang X.-S. Firefly algorithms for multimodal optimization, *Stochastic Algorithms: Foundations and Applications*, 2009, Vol. 5792, pp. 169–178. DOI: 10.1007/978-3-642-04944-6_14
6. Breiman, L. Random forests, *Machine Learning*, 2001, No. 1, pp. 5–32. DOI: 10.1023/A:1010933404324
7. Random forests – Classification manual [Electronic resource], 2017, Access mode: http://www.math.usu.edu/adele/Forests/
8. Strobl C., Zeileis A. Danger: High power! – exploring the statistical properties of a test for random forest variable importance, *Proceedings in Computational Statistics*, 2008, pp. 59–66.
9. Xue B., Zhang M., Browne W. N. Particle swarm optimization for feature selection in classification: Novel initialization and updating mechanisms, *Applied Soft Computing*, 2014, Vol. 18, pp. 261–276. DOI: 10.1109/TSMCB.2012.2227469
10. Feng D., Chen F., Xu W. Supervised feature subset selection with ordinal optimization, *Knowledge-Based Systems*, 2014, Vol. 56, pp. 123–140. DOI: 10.1016/j.knosys.2013.11.004
11. Bouaguel W., Mufti G. B., Limam M. A fusion approach based on wrapper and filter feature selection methods using majority vote and feature weighting, *Proceedings of the International Conference on Computer Applications Technology*, 2013, pp. 1–6. DOI: 10.1109/ICCAT.2013.6522003
12. Wang G., Ma J., Yang S. An improved boosting based on feature selection for corporate bankruptcy prediction, *Expert Systems with Applications*, 2014, Vol. 41, No. 5, pp. 2353–2361. DOI: 10.1016/j.eswa.2013.09.033
13. Srinivasa K. G. Application of Genetic Algorithms for Detecting Anomaly in Network Intrusion Detection Systems, *Advances in Computer Science and Information Technology. Networks and Communications*, 2012, Vol. 84, pp. 582–591. DOI: 10.1007/978-3-642-27299-8_61
14. Yu K. M., Wu M. F., Wong W. T. Protocol-based classification for intrusion detection, *Applied Computer and Applied Computational Science*, 2008, Vol. 3, No. 3, pp. 135–141.
15. Akbar S., Nageswara R. K., Chandulal J. A. Intrusion detection system methodologies based on data analysis, *International Journal of Computer Applications*, 2010, Vol. 5, No. 2, pp. 10–20. DOI: 10.5120/892-1266
16. Sethuramalingam S., Naganathan E. R. Hybrid feature selection for network intrusion detection, *International Journal of Computer Science and Engineering*, 2011, Vol. 3, No. 5, pp. 1773–1780. DOI: 10.4225/75/57a84d4fbefbb
17. Banati H., Bajaj M. Fire Fly based feature selection approach, *IJCSI International Journal of Computer Science Issues*, 2011, Vol. 8, № 4, pp. 473–80.
18. Hothorn T., Hornik K., Zeileis A. Unbiased recursive partitioning: a conditional inference framework, *Journal of Computational and Graphical Statistics*, 2006, Vol. 15, No. 3, pp. 651–674. DOI: 10.1198/106186006X133933
19. Breiman L. Stacked Regressions, *Machine Learning*, 1996, Vol. 24, pp. 49–64. DOI: 10.1007/BF00117832
20. Strobl C., Boulesteix A.-L., Kneib T., Augustin T., Zeileis A. Conditional variable importance for random forests, *BMC Bioinformatics*, 2008, Vol. 9, No. 1, pp. 25. DOI: 10.1186/1471-2105-9-307
21. Siroky D. Navigating Random Forests and related advances in algorithmic modeling, *Statistics Surveys*, 2009, Vol. 3, pp. 147–163. DOI: 10.1214/07-SS033
22. Archer K. J., Kimes R. V. Empirical characterization of random forest variable importance measures, *Computational Statistics & Data Analysis*, 2008, No. 4, pp. 2249–2260. DOI: 10.1016/j.csda.2007.08.015
23. Strobl C., Boulesteix A.-L., Zeileis A., Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution, *BMC Bioinformatics*, 2007, Vol. 8, No. 1, pp. 1471–2105. DOI: 10.1186/1471-2105-8-25
24. Liaw A., Wiener M. Classiыcation and Regression by randomForest. R News, 2002, Vol. 2, No. 3, pp. 18–22.
25. Aggarwal P., Sharma S. K. Analysis of KDD dataset attributes-class wise for intrusion detection, *Procedia Computer Science*, 2015, Vol. 57, pp. 842–851. DOI: 10.1016/j.procs.2015.07.490
26. McHugh J. Testing intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by lincoln laboratory, *ACM Transactions on Information and System Security*, 2000, Vol. 3, No. 4, pp. 262–294. DOI: 10.1145/382912.382923
27. Tavallaee M., Bagheri E., Lu W., Ghorbani A. A detailed analysis of the KDD CUP 99 Data Set, *Proceedings of the second IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 2009, pp. 53–58. DOI: 10.1109/CISDA.2009.5356528
28. NSL-KDD data set for network-based intrusion detection systems [Electronic resource], 2017, Access mode: http://nsl.cs.unb.ca/NSL-KDD/
29. Davis J. J., Clark A. J. Data preprocessing for anomaly based network intrusion detection: A review, *Computers & Security*, 2011, Vol. 30, No. 6–7, pp. 353–375. DOI: 10.1016/j.cose.2011.05.008
30. Holz T. 13 security measurements and metrics for networks, *Dependability Metrics*, 2008, pp. 157–165. DOI: 10.1007/978-3-540-68947-8_13

УДК 378.147:044.4'24(477)

Кузьма К. Т.

Канд. техн. наук, старший викладач кафедри комп'ютерної інженерії, Миколаївський національний університет ім. В. О. Сухомлинського, Миколаїв, Україна

ОБЧИСЛЮВАЛЬНА ТЕХНОЛОГІЯ ПЕРЕВІРКИ РІВНЯ ЗНАНЬ НА ОСНОВІ МЕТОДУ ПОСЛІДОВНОГО АНАЛІЗУ

Актуальність. Вирішено актуальну задачу підвищення ефективності процесу підтримки прийняття рішень під час статистичного контролю знань.

Мета роботи – розробка обчислювальної процедури вирішення чотирьохальтернативної задачі класифікації тестованих за рівнем навченості, що дозволяє здійснювати контроль знань диференційовано, мінімізує об'єм завдань, необхідний для виконання.

Метод. Запропоновано обчислювальну процедуру класифікації тестованих на чотири класи, які відповідають рівням навченості: «початковий», «середній», «достатній», «високий», що базується на використанні двохальтернативного критерію послідовного аналізу в декілька етапів та забезпечує здійснення контролю знань в процесі виконання завдань, мінімізуючи таким чином час перевірки знань, що дозволяє автоматизувати процес перевірки статистичних гіпотез у системах тестування та навчання з метою диференційної оцінки знань учасників навчального процесу. Для вирішення задачі оцінки придатності тесту запропоновано метод, що базується на побудові функції оперативної характеристики послідовного критерію, яка дозволяє визначити обсяг завдань достатній для досягнення бажаного рівня якості тесту за рахунок встановлення зв'язку між очікуваною ймовірністю прийняття гіпотези та випадковим значенням параметра ймовірності появи у виборці з 1,2,…n питань приймального числа невірно виконаних завдань.

Результати. Розроблено програмне забезпечення, яке реалізує запропоновану обчислювальну процедуру, що використано при проведенні обчислювальних експериментів тестового контролю знань.

Висновки. Проведені експерименти підтвердили працездатність запропонованої процедури і програмного забезпечення, що її реалізує, а також дозволяють рекомендувати їх для застосування на практиці для рішення задач автоматизованої перевірки рівня знань.

Ключові слова: послідовний аналіз, перевірка рівня знань, класифікація тестованих за рівнем знань, перевірка гіпотез, критеріально-орієнтований тест.

НОМЕНКЛАТУРА

ОТ – обчислювальна технологія;

СППР – система підтримки прийняття рішень;

d_h – число невірно виконаних завдань серед h перевірених;

h – кількості питань у виборці;

H_i – гіпотеза щодо віднесення об'єкта навчання до i-го класу навченості;

$L(p)$ – оперативна характеристика статистичного критерію;

n – загальна кількість тестових питань;

p_i – ймовірність невиконання завдань для i рівня навчальних досягнень, $i=\overline{1,4}$;

$f(d)$ – функція щільності розподілу неправильних відповідей;

$f(Q)$ – функція оцінювання знань;

μ_1 – математичне очікування числа правильно виконаних завдань для неатестованих об'єктів навчання, що відповідає гіпотезі A_1 – об'єкт навчання не володіє достатніми знаннями та не атестується;

μ_2 – математичне очікування числа правильно виконаних завдань для атестованих об'єктів навчання, що відповідає гіпотезі A_2 – об'єкт навчання володіє достатніми знаннями та атестується позитивно;

σ^2 – дисперсія випадкової величини X коректно виконаних завдань для атестованих та неатестованих об'єктів навчання;

X_0 – допустиме (прийнятне) число коректно виконаних завдань;

p_{a_i} – приймальна доля невірно виконаних завдань для i-ї класифікації;

p_{r_i} – неприймальна доля невірно виконаних завдань для i-ї класифікації;

g_i – i-та група із h завдань;

a_h – прийнятне число невірно виконаних завдань для певного класу;

r_h – неприйнятне число невірно виконаних завдань для певного класу;

$a_i(h)$ – пороги прийняття гіпотези щодо зарахування об'єкта навчання до одного з чотирьох класів навченості, $i=\overline{1,3}$;

$r_i(h)$ – пороги неприйняття гіпотези щодо зарахування об'єкта навчання до одного з чотирьох класів навченості, $i=\overline{1,3}$;

m – номер питання;

L – вага відповіді: $L=0$ – вірна відповідь, $L=1$ – невірна відповідь;

q – стандарт оцінювання;

$n_{\text{ср}_{\max}}$ – максимальний обсяг завдань, відповіді на які необхідно перевірити;

$P_n(d)$ – ймовірність появи d невірних відповідей з перевірених n-завдань;

DOI 10.15588/1607-3274-2017-3-14

α – ймовірність помилки 1-го роду, ймовірність неприйняття головної гіпотези, якщо доля невірних відповідей менша за критеріальну;

β – ймовірність помилки 2-го роду, ймовірність прийняття головної гіпотези, якщо доля невірних відповідей більша за критеріальну;

C – приймальне число невірних відповідей для заданої класифікації;

C_{ij} – ваги прийняття рішень, $i=\overline{1,2}$; $j=\overline{1,2}$.

ВСТУП

Тест, як інструмент стандартизованої процедури проведення і заздалегідь спроектованої технології обробки та аналізу результатів, є зручним засобом вимірювання навчальних досягнень. Тестування як форма контролю навчальних досягнень студентів в останні роки все частіше стала застосовуватися в системі вищої, професійної освіти, оскільки є технологією, що дозволяє об'єктивно і швидко оцінити рівень досягнень кожного студента.

За цілями застосування педагогічні тести поділяються на два класи – нормативно-орієнтовані і критеріально-орієнтовані. Нормативно-орієнтований тест (norm-referenced test) дозволяє ранжувати випробовуваних за рівнем знань. Критеріально-орієнтований тест (criterion-referenced test) дозволяє виявити рівень засвоєння випробовуваним певного розділу в заданій предметній галузі. Зазвичай тестовий бал відображає частку правильно виконаних завдань і виражається у відсотках. При використанні критеріально-орієнтованого підходу особлива увага приділяється методиці оптимального вибору критеріального балу (або балів). Саме при критеріально-орієнтованому тестуванні задача перевірки знань розглядається як задача підтримки прийняття рішень та класифікації тестованих за рівнем їх підготовки.

Рішення цієї задачі забезпечує необхідну об'єктивність при оцінці та контролі знань, можливість мінімізації об'єму завдань, який повинен вирішити опитуваний. Метою роботи є розробка обчислювальної процедури вирішення чотирьохальтернативної задачі класифікації тестованих за рівнем навченості, що дозволяє здійснювати контроль знань диференційовано, мінімізує об'єм завдань, необхідний для виконання.

1 ПОСТАНОВКА ЗАДАЧІ

На основі проведених досліджень, результати яких наведено в роботі [1], для вирішення задачі прийняття рішень щодо класифікації об'єктів навчання під час тестування обрано метод послідовного аналізу, оскільки його застосування дозволяє підвищити ефективність процесу перевірки знань за рахунок мінімізації часових витрат на його проведення.

На сьогодні для процесу ідентифікації рівня знань учасників навчального процесу з використанням методу послідовного аналізу розроблено обчислювальні процедури перевірки двохальтернативних випадків, наприклад, класифікації об'єктів навчання на дві групи: атестовані та не атестовані. Постає задача розробки алгоритму класифікації тестованих відносно декількох гіпотез. Правилом прийняття рішень є формування границь, за допомогою яких здійснюється класифікація тестованих на чотири групи (I група – «високий» рівень навченості, II – «достатній», III – «середній», IV – «початковий»).

Дано відповіді $(x_1,...,x_h)$ на h запитань із загальної кількості завдань n. Припустимо, що $x_i=0$, якщо завдання виконано вірно, $x_i=1$, якщо завдання виконано невірно. Нехай p_i визначає відносне число невиконаних завдань. Для процесу перевірки відповідей на тестові питання під час контролю знань постає задача перевірки гіпотези про те, що p_i на деякому діапазоні h виконаних завдань із загальної кількості завдань не перевищує деякої заданої величини p_i'. Таким чином, задачу оцінки навчальних досягнень необхідно вирішити шляхом послідовної перевірки гіпотез $p_i \le p_i'$ проти гіпотез $p_i > p_i'$, де $i=\overline{1,4}$ – номер класу навченості.

2 ОГЛЯД ЛІТЕРАТУРИ

Задача обробки даних педагогічного тестування та контролю знань за критеріально-орієнтованою методикою в роботах [2–5] розглядається як задача підтримки прийняття рішень при класифікації об'єктів навчання (тестованих) за рівнем підготовки, для вирішення якої використовуються методи теорії статистичних рішень (критерій Байєса, Неймана-Пірсона, мінімакса, Вальда). Застосування даних методів спрямоване на створення систем статистичного контролю знань за альтернативною ознакою.

При тестуванні за альтернативною ознакою використовується закрита форма тесту, характеристиками якої є: функція щільності розподілу неправильних відповідей – $f(d)$, прийнятний рівень неправильних відповідей – p_0, неприйнятний рівень неправильних відповідей – p_1, ризик «заниженої» оцінки знань – α, ризик «завищеної» оцінки знань – β, функція оцінювання знань – $f(Q)$, обсяг освітньої інформації – N, обсяг вибірки завдань тесту – n та критерій прийняття розв'язків у вигляді граничного числа неправильних відповідей – X.

У роботі [2] запропоновано правило прийняття рішень, яке визначає допустиме (прийнятне) число коректно виконаних завдань X_0 (границя прийняття рішень) на основі критерію Байєса та поділяє всіх тестованих на дві групи: атестовані та неатестовані. Умовні щільності ймовірності випадкової величини X коректно виконаних завдань для атестованих та неатестованих об'єктів навчання подаються у вигляді гаусовських законів розподілення:

$$f\left(X/\mu_i\right)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{\left(X-\mu_i\right)^2}{2\sigma^2}\right\},\ i=1,2.$$

Стратегія прийняття рішень відносно границі X_0 має чотири результати: вірна гіпотеза A_1 та приймається рішення про A_1; вірна гіпотеза A_1, а приймається рішення про A_2; вірна гіпотеза A_2 та приймається рішення про A_2; вірна гіпотеза A_2, а приймається рішення про A_1.

Перший та третій результат є правильними рішеннями, другий та четвертий – помилковими.

Байєсовський критерій базується на двох припущеннях: апріорні ймовірності гіпотез P_1 та P_2 задані та $P_1+P_2=1$; задані ваги чотирьох зазначених дій: C_{11}, C_{12}, C_{22}, C_{21}. Слідуючи правилам прийняття рішень для вибору або A_1, або A_2, простір спостережень N ділиться на дві

частини N_1 та N_2. Якщо результат тестування потрапляє в N_1, то приймається гіпотеза A_1, якщо в $N_2 - A_2$.

Байєсовський критерій будує рішення так, щоб у середньому втрати та ризики були мінімальними. Значення величини очікуваного ризику подається виразом:

$$R = P_1C_{21} + P_2C_{22} + \int_{N_1} \{[P_2(C_{12} - C_{22}) f_2(X/\mu_2)] -$$

$$-[P_1(C_{21} - C_{11}) f_1(X/\mu_1)]\} dx.$$

Щоб мінімізувати ризик R за умови, що другий член підінтегрального виразу більший, ніж перший (значення вірних відповідей менше значення невірних), необхідно всі значення X включати до N_1 та навпаки. Таким чином, області рішень N_1 та N_2, які відповідають гіпотезам A_1 та A_2 визначаються за наступних умов:

$$\text{якщо } P_2(C_{12} - C_{22}) f_2\left(X/\mu_2\right) \geq P_1(C_{21} - C_{11}) f_1\left(X/\mu_1\right), \quad (1)$$

то X відноситься до N_2 – приймається рішення щодо істинності гіпотези A_2, та навпаки.

Вираз 1 подається у вигляді нерівності двох відношень:

$$\frac{f_2(X/\mu_2)^{A_2}}{f_1(X/\mu_1)} > \frac{P_1(C_{21} - C_{11})}{{}_{A_1}P_2(C_{12} - C_{22})}.$$

Ліва частина називається відношенням правдоподібності і визначається $\Lambda(x)$. Права частина називається границею прийняття рішення й визначається η та є const, оскільки залежить від апріорних параметрів задачі. Таким чином, Байєсовський критерій зводиться до критерію відношення правдоподібності, записується у вигляді нерівності $\Lambda(x)^{A_2} >_{A_1} \eta$ та трактується наступним чином: якщо величина відношення правдоподібності для двох гіпотез більше границі прийняття рішень, то приймається гіпотеза A_2, якщо менше, то – A_1.

Мінімаксний критерій використовується при відсутності інформації щодо значення апріорних ймовірностей гіпотез A_1 та $A_2 - P_1$ та P_2 відповідно. Якщо не визначені апріорні ймовірності гіпотез P_1 та P_2 і не задані ваги прийняття рішень C_{ij}, то для рішення задачі використовують критерій Неймана-Пірсона та значення ймовірностей помилок першого та другого роду.

Алгоритми прийняття рішень на основі критеріїв Байєса, Неймана-Пірсона та мінімаксу щодо класифікації об'єктів навчання передбачають, що N завдань, які видаються для об'єктів навчання, та X – число вірно виконаних завдань фіксовані. Якщо кількість вірно виконаних завдань не фіксовано, то прийняття рішень щодо класифікації тестованих, не перевіряючи всіх завдань, здійснюється на основі послідовного аналізу відповідей з використанням критерію Вальда.

3 МАТЕРІАЛИ ТА МЕТОДИ

На основі методів, поданих у роботах [2, 6, 7], розроблено алгоритм для чотирьохальтернативної задачі класифікації учасників навчального процесу. Пропонується класифікувати об'єкти навчання за рівнем знань: I клас – об'єкти навчання, які володіють «високим» рівнем знань (A, від 90 до 100 балів); II клас – «достатнім» (BC від 65 до 89 балів); III клас – «середнім» (DE від 50 до 64 балів); IV клас – «початковим» (F, FX від 1 до 49 балів).

Сутність класифікації полягає в розбиті інтервалу тестових завдань N на чотири області: N_1, N_2, N_3, N_4. Якщо число невірно виконаних завдань d_h потрапило в область N_1, то приймається гіпотеза H_1 – об'єкт навчання має «високий» рівень знань; якщо в N_2, то приймається гіпотеза H_2 (відповідає «достатньому» рівню знань); якщо в N_3 – гіпотеза H_3 (відповідає «середньому» рівню знань); інакше – приймається гіпотеза H_4 (відповідає «початковому» рівню знань).

Розроблений алгоритм базується на використанні двохальтернативного критерію послідовного аналізу в декілька етапів. Якщо p_i визначає відносне число невиконаних завдань на деякому діапазоні h, то ймовірність отримання вибірки $(x_i,...,x_h)$ обчислюється за формулою:

$$p_h = p^{d_h}(1-p)^{h-d_h}. \quad (3)$$

На першому етапі рішення задачі визначимо гіпотези оцінки рівня навчальних досягнень тестованих. Оскільки ймовірність невиконання завдань p_i є випадковою величиною, то приймаємо наступні гіпотези: H_1: якщо ймовірність невиконання завдання дорівнює p_1 $(p = p_1)$ і тестований зараховується до класу I; H_2: якщо ймовірність невиконання завдання дорівнює p_2 $(p = p_2)$ і тестований зараховується до класу II; H_3: якщо ймовірність невиконання завдання дорівнює p_3 $(p = p_3)$ і тестований зараховується до класу III; H_4: якщо ймовірність невиконання завдання дорівнює p_4 $(p = p_3)$ і тестований зараховується до класу IV.

Допустима доля невірних відповідей для кожного класу p_i встановлюється на основі попередніх досліджень перевірки знань у контрольній групі або виходячи з ефективності контролю. При цьому перевірка здійснюється наступним чином. Із загальної кількості n завдань вибирається група g_1, яка містить h завдань.

Перевірка закінчується прийняттям гіпотези про атестацію тестованого, якщо в групі g_1 для відповідного класу виконується умова: $a_h \geq d_h$. Перевірка закінчується прийняттям гіпотези про неатестацію тестованого, якщо в групі g_1 для відповідного класу виконується умова: $r_h \leq d_h$. Якщо $a_h < d_h < r_h$, обирається друга група g_2, яка містить наступні h завдань. Знову тестований атестується позитивно, якщо загальне число невиконаних завдань у двох групах d_{2h} менше або дорівнює a_{2h}; тестований неатестується, якщо $d_{2h} \geq r_{2h}$ та береться третя група g_3 із h запитань, якщо $a_{2h} < d_{2h} < r_{2h}$. Процес продовжується, поки тестований буде атестований або неатестований. Таким чином, коли спостереження проводяться над групами по h завдань, число визначених невірно виконаних завдань d_m порівнюється з відповідним приймальним числом a_m, або неприймальним числом r_m тільки при $m = h, 2h, 3h,...n$.

Блок-схему ОТ перевірки рівня знань на основі методу послідовного аналізу наведено на рис. 1.

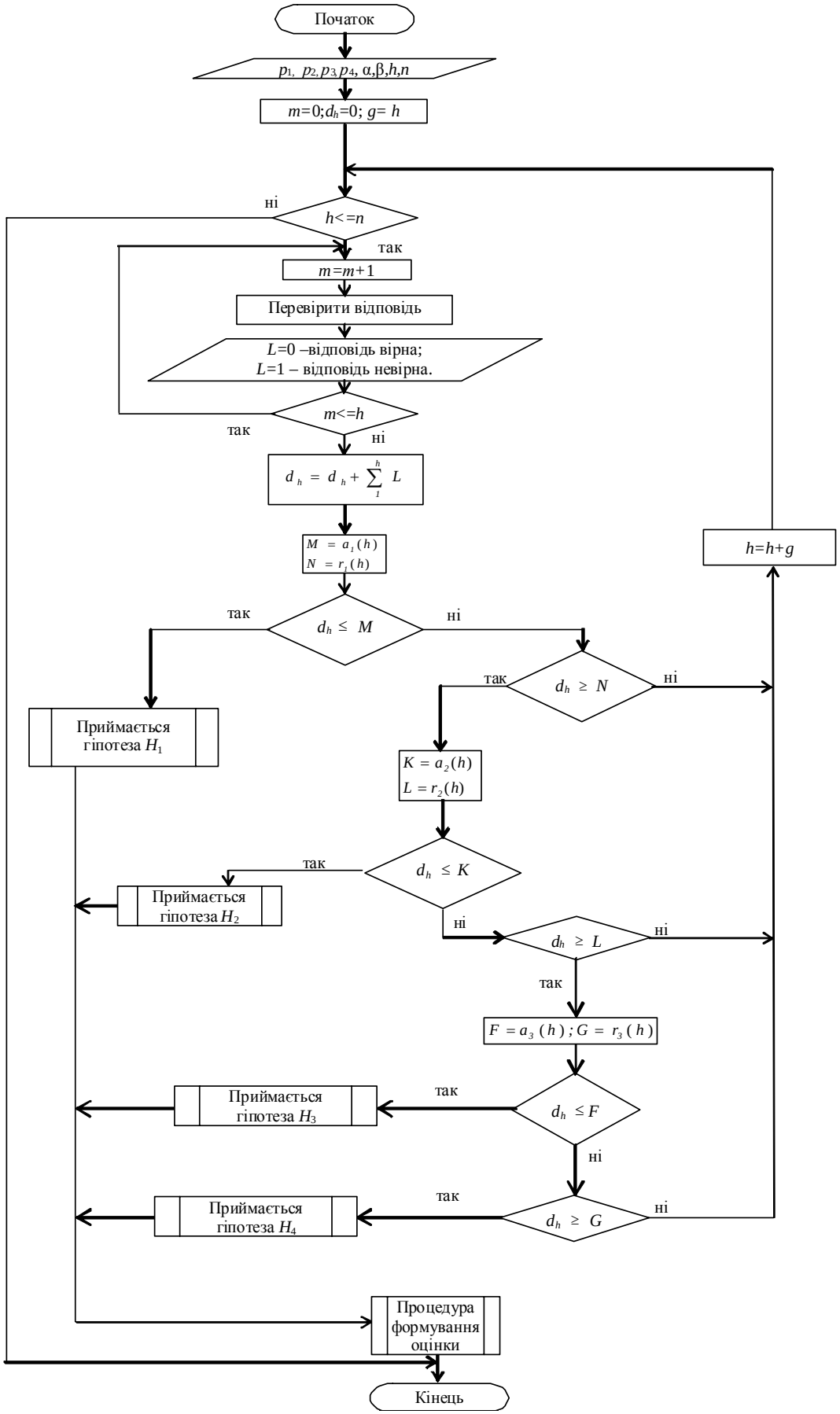


Рисунок 1 – Блок-схема алгоритму ОТ перевірки рівня знань на основі методу послідовного аналізу

Алгоритм ОТ передбачає виконання наступних етапів:

Етап 1. Уведення меж розбивки вихідного параметра – ймовірностей $p_i, i=\overline{1,4}$ невиконання завдань для кожного рівня навчальних досягнень.

Етап 2. Введення α, β, n, h.

Етап 3. Цикл за всіма діапазонами. Визначається на поточному діапазоні виконаних завдань h значення вихідного параметру d_h. Послідовно в поточному діапазоні обчислюються пороги прийняття $a_i(h), i=\overline{1,3}$ або неприйняття $r_i(h), i=\overline{1,3}$ гіпотези щодо зарахування об'єкта навчання до одного з чотирьох класів навченості за формулами:

$$a_i(h)=\frac{\ln\frac{\beta}{1-\alpha}}{\ln\frac{p_{r_i}}{p_{a_i}}-\ln\frac{1-p_{r_i}}{1-p_{a_i}}}+\frac{h\cdot\ln\frac{1-p_{a_i}}{1-p_{ri}}}{\ln\frac{p_{r_i}}{p_{a_i}}-\ln\frac{1-p_{r_i}}{1-p_{a_i}}};$$

$$r_i(h)=\frac{\ln\frac{1-\beta}{\alpha}}{\ln\frac{p_{r_i}}{p_{a_i}}-\ln\frac{1-p_{r_i}}{1-p_{a_i}}}+\frac{h\cdot\ln\frac{1-p_{a_i}}{1-p_{ri}}}{\ln\frac{p_{r_i}}{p_{a_i}}-\ln\frac{1-p_{r_i}}{1-p_{a_i}}}\ i=\overline{1,3}.$$

Етап 4. Перевіряється, в яку область попадає значення d_h. Якщо значення d_h не попадає в границі відповідної області, здійснюється перехід до наступного діапазону питань $g_i\ (x_{h+1},...,x_{ih})$ де $i=2,...$. Якщо $h=n$, а рішення ще не прийнято, відбувається усічення послідовного критерію з використанням наступного правила: у випадку коли $d_h \ge (a_i(h)+r_i(h))/2$ приймається рішення щодо перевірки гіпотез $H_{i+1},...,H_4$, а при $d_h < (a_i(h)+r_i(h))/2$, приймається гіпотеза H_i, тестування завершується.

Для визначення ймовірності невиконання завдань p_i, $i=\overline{1,4}$ необхідно здійснити вибір стандарту оцінювання q_i. Методи вибору тестового стандарту засновані на експертних оцінках змісту тестового завдання. На основі дослідження даних методів, описаних у роботах [8–10], обрано метод Ангофф, який заснований на послідовних експертних оцінках змісту тестових завдань. Спочатку виконується вибір стандарту оцінювання для IV класу. Експерт-викладач для кожного завдання тесту встановлює ймовірність того, що мінімально компетентний студент дасть на нього вірну відповідь. Для однозначності експерту (або групі експертів) пропонується обрати значення ймовірності P_i зі значень 0,9, 0,8 ,..., 0,1. Визначивши суму значень даних ймовірностей отримаємо критеріальний бал: $K=\sum_{i=1}^{n} P_i$. Стандарт оцінювання визначається за формулою: $q=\frac{K}{n}$.

Тоді ймовірність невиконання завдань $p=1-q$. Після вибору критеріального балу для «початкового» рівня навченості, експерт проводить оцінку кожного тестового завдання вже на більш високий стандарт «середній», «достатній» та «високий».

Процедура контролю знань потребує вирішення задачі оцінки придатності тесту, що полягає у виборі таких критеріїв класифікації (α, β, p_i, $i=\overline{1,4}$), які б зробили помилки першого та другого роду малоймовірними та забезпечили визначення обсягу завдань достатнього для досягнення бажаного рівня якості тесту (ймовірність віднесення об'єктів навчання, які мають індивідуальний бал вищий або нижчий критеріального на величину не більше 10%, до певного класу навченості не повинна бути меншою за 0,8). Для розв'язку даної задачі використовується метод, що базується на побудові функції оперативної характеристики послідовного критерію відношення ймовірностей, яка дозволяє встановити зв'язок між очікуваною ймовірністю прийняття гіпотези та випадковим значенням параметра ймовірності появи у виборці з 1,2,… питань приймального числа невірно виконаних завдань:

$$L(p)=P_0+P_1+...+P_C=\sum_{d=0}^{C}P_n(d). \qquad (4)$$

Якість обраних критеріїв класифікації визначається рівняннями: $L(p)\ge 1-\alpha$, якщо $p=p_{\alpha_i}$; $L(p)\le\beta$, якщо $p=p_{r_i}$. Значення оперативної характеристики статистичного критерію для кожного класу обчислюється за формулою:

$$L(p)\approx([(1-\beta)/\alpha]^l-1)/([(1-\beta)/\alpha]^l-[\beta/(1-\alpha)]^l),$$

де параметр l змінюється від $-\infty$ до $+\infty$ та визначається із рівняння:

$$p_i=\left[1-\left((1-p_{r_i})/(1-p_{a_i})\right)^l\right]\Big/\left[\left(p_{r_i}/p_{a_i}\right)^l-\left((1-p_{r_i})/(1-p_{a_i})\right)^l\right],\ i=\overline{1,3}.$$

Максимальний обсяг завдань, відповіді на які необхідно перевірити для прийняття рішень щодо наявності відповідної гіпотези H_1, H_2, H_3, пропонується обчислювати за формулою:

$$n_{\text{ср}_{\max}}=\sum_{i=1}^{3}\left[-\left(\ln\frac{\beta}{1-\alpha}\right)\cdot\left(\ln\frac{1-\beta}{\alpha}\right)\right]\Big/\left[3\cdot\ln\frac{p_{r_i}}{p_{a_i}}\cdot\ln\frac{1-p_{a_i}}{1-p_{r_i}}\right].$$

4 ЕКСПЕРИМЕНТИ

Для проведення експерименту необхідно виділити контрольну групу студентів та з використанням СППР «ManageEdu» [11] або власноруч розробленого програмного продукту, який реалізовуватиме запропонований алгоритм класифікації тестованих, провести підсумковий контроль знань з будь-якої дисципліни за традиційною методикою та з використанням запропонованої обчислювальної технології перевірки рівня знань на ос-

нові методу послідовного аналізу. Програмний додаток повинен реалізовувати наступні функції: розраховувати ймовірнісні характеристики можливого віднесення тестованого до того чи іншого класу навченості (p_1, p_2, p_3, p_4, α, β, n, h), формувати границі прийняття рішень для чотирьох гіпотез, для перевірки якості обраних параметрів будувати оперативні характеристики послідовного критерію для гіпотез H_1, H_2, H_3, проведення контролю знань з використанням запропонованої ОТ та стандартним методом.

5 РЕЗУЛЬТАТИ

Для проведення експерименту було задіяно контрольну групу студентів (23 чоловіка) ВНЗ «Міжнародний технологічний університет «Миколаївська політехніка» III курсу економічного факультету з дисципліни «Фінансовий облік».

З використанням модуля «Последовательный анализ» СППР «ManageEdu» були сформовані ймовірнісні характеристики можливого віднесення тестованого до того чи іншого класу навченості: p_1=0,1; p_2=0,35 ; p_3=0,6 ; p_4=0,75; $\alpha = 0{,}05$; $\beta = 0{,}01$; n =100; h =10.

Значення параметрів границь прийняття рішень, сформовані на основі запропонованого алгоритму, наведено на рис. 2. в табличному та графічному вигляді.

Для перевірки якості обраних параметрів побудовано оперативні характеристики послідовного критерію для гіпотез H_1, H_2, H_3. Аналіз графіка оперативної характеристики $L(p)$ для гіпотези H_1 підтверджує оптимальність обраних α,β, приймального числа p_1 та неприймального числа p_2: очікуваний відсоток студентів, які будуть атестовані «на відмінно» при 10–16% невірних відповідей з масиву тестових завдань складає 80%, $L(p) > 80$ %. Оптимальна кількість завдань для даної класифікації – 24.

Для даного тесту середній обсяг завдань, який є достатнім для досягнення бажаного рівня якості процедури тестування, складає $n_{ср} = \dfrac{(n_{срI} + n_{срII} + n_{срIII})}{3} = 45.$

Якщо при постійній кількості завдань збільшувати приймальне число відповідей, то контроль буде менш «жорстким» – збільшуються ймовірності атестації об'єктів навчання за відповідними рівнями навченості. Та навпаки: якщо при постійному значенні приймального числа збільшувати кількість завдань, – контроль буде більш «жорстким».

Після збереження отриманих значень вихідних параметрів процедури послідовного аналізу в діалоговому вікні налаштування параметрів теми на стороні клієнтської частини програмного комплексу відбувається процес контролю знань.

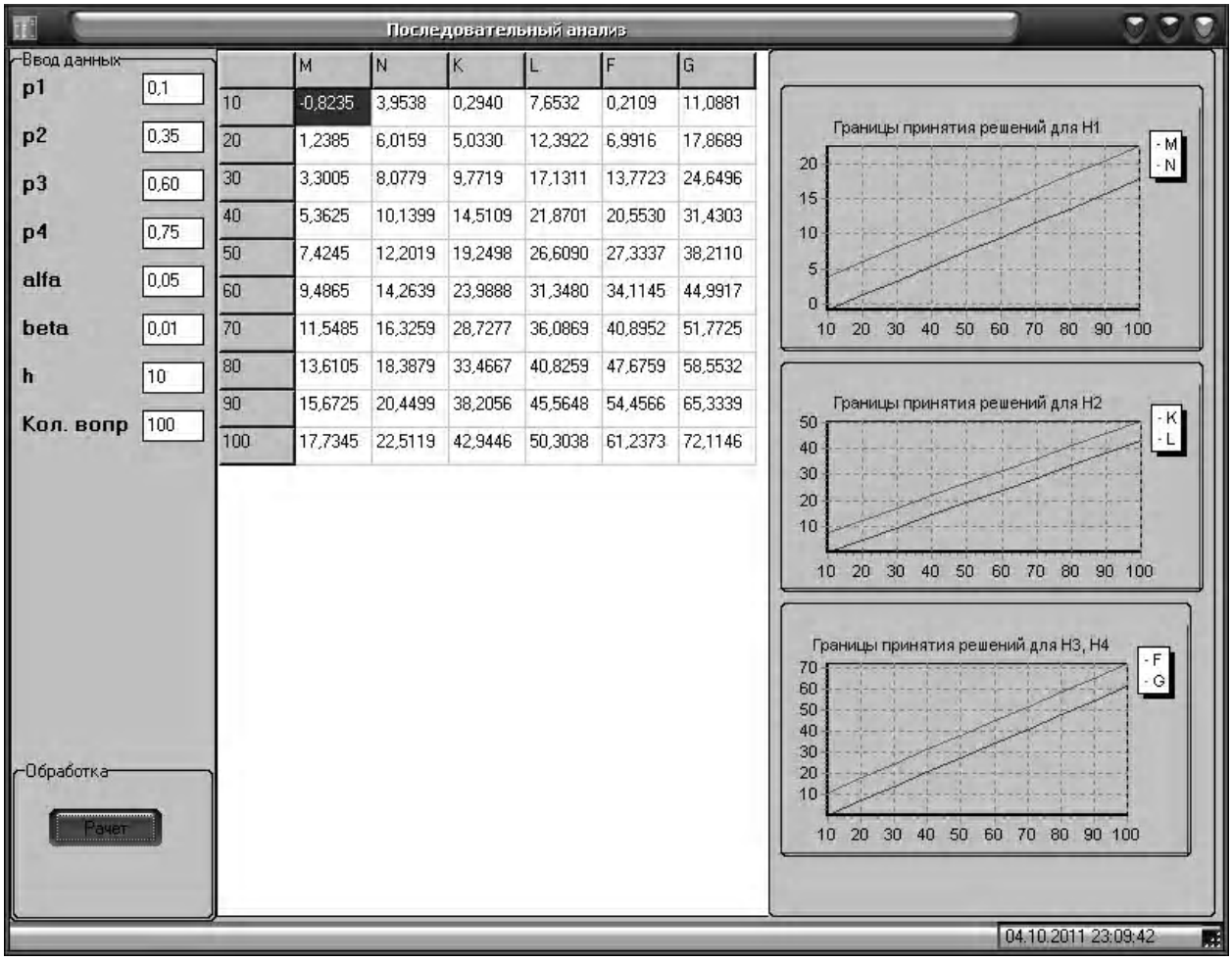


Рисунок 2 – Діалогове вікно формування границь прийняття рішень

На рис. 3. наведено приклад вирішення задачі тестування на основі процедури послідовного аналізу для одного об'єкта навчання.

Таким чином, якщо тестований з десяти завдань отриманих на першому етапі тестування виконав вірно підряд вісім, то згідно з правилами прийняття рішень йому буде видана наступна вибірка з десяти питань та продовжена перевірка гіпотези H_1. Після перевірки гіпотези H_1, буде прийняте рішення про перевірку гіпотези H_2, відповідно до якої студент отримає наступні десять завдань, оскільки $K < d_h < L$. Після виконання 30 завдань приймається рішення щодо прийняття гіпотези H_2 при $d_h = 7$ та зарахування студента до класу II, оскільки $d_h \le K$. Тестований отримає оцінку «BC» за системою ECTS або «4» за 5-ти бальною системою.

6 ОБГОВОРЕННЯ

В порівнянні з двохальтернативною класифікацією тестованих на дві групи: «атестовані» та «неатестовані», представленою в роботі [1], запропонована чотирьохальтернативна класифікація дозволяє здійснювати оцінку навчальних досягнень диференційовано, є адаптованою для автоматизованого контролю знань у вищих закладах освіти.

На рис. 4. зображено розподіл студентів за класами навченості за результатами перевірки знань з використанням ОТ послідовного аналізу та стандартного режимів тестування.

Оцінка якості класифікації виконується шляхом визначення наскільки тісно розташовані об'єкти в класах у порівнянні з розташуванням об'єктів у всій групі для стандартного та адаптивного режимів перевірки. Для IV класу («початковий» рівень навченості) відхилення результатів «адаптивного» режиму тестування від стандартного складає +4%, для III – – 4%, для II та I – 0%. Отже, класифікації рівнозначні, при цьому для стандартного режиму тестування середня кількість завдань, яку виконує кожний студент дорівнює «50», а для адаптивного – «30».

Запропонована обчислювальна технологія дає можливість системі тестування приймати рішення щодо класифікації поведінки студента, не перевіряючи всі n завдань, що скорочує число завдань, які необхідно перевірити, та забезпечує індивідуальну мінімізацію часу навчання.

Додатковими умовами забезпечення високої ефективності застосування запропонованої технології контролю знань є:

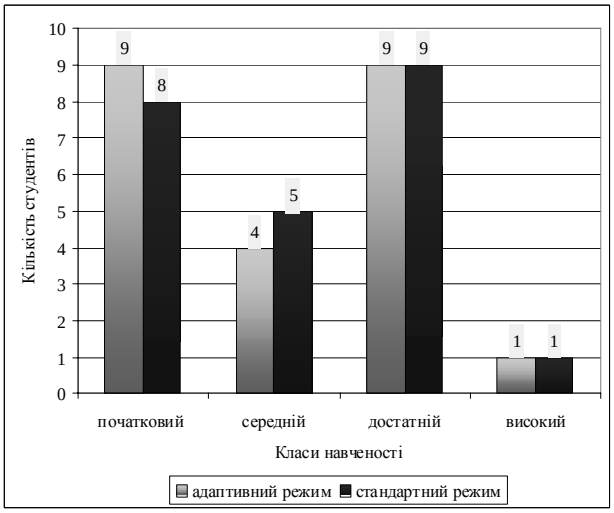
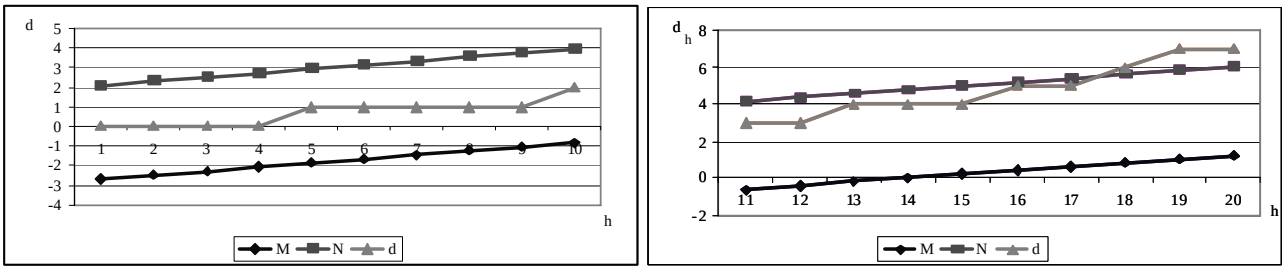
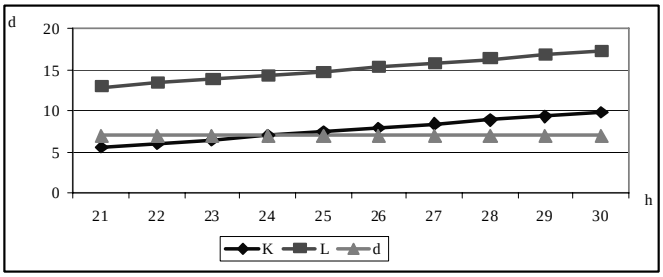


Рисунок 4 – Розподіл студентів за класами навченості на основі ОТ послідовного аналізу та стандартного режиму тестування



а б



в

$$d_h = (0\,0\,0\,0\,1\,1\,1\,1\,1\,2\,3\,3\,4\,4\,4\,5\,5\,6\,7\,7\,7\,7\,7\,7\,7\,7\,7\,7)^{\mathrm{T}}$$

$$h = (1\,2\,3\,4\,5\,6\,7\,8\,9\,10\,11\,12\,13\,14\,15\,16\,17\,18\,19\,20\,21\,22\,23\,24\,25\,26\,27\,28\,29,\,30)^{\mathrm{T}}$$

Рисунок 3 – Графіки границь прийняття рішень:

а – відносно гіпотези H_1 при $h = 10$; б – відносно гіпотези H_1 при $h = 20$; в – відносно гіпотези H_2 при $h = 30$

1. Вимога підвищення якості тестових завдань, основними показниками якої є валідність і надійність тестів. Дані показники характеризують, відповідно, стабільність отриманих результатів тестування і їх здатність правильно відображати рівень підготовки.

2. Можливість у процесі тестування не тільки визначення рівня навченості слухачів, але й швидкого відновлення знань шляхом перегляду правильної відповіді і отримання (при необхідності) іншої додаткової дидактичної інформації з предмета.

ВИСНОВКИ

В роботі розглянуто задачу оцінки знань тестованих як задачу їх класифікації за рівнем знань, умінь та навичок, для рішення якої використано послідовну процедуру перевірки гіпотез і правила прийняття рішень на основі критерію Вальда. Розроблено обчислювальну процедуру класифікації тестованих на чотири класи, які відповідають рівням навченості: «початковий», «середній», «достатній», «високий».

Проаналізовані обчислювальні технології підтримки прийняття рішень під час контролю знань об'єктів навчання, які базуються на методах теорії статистичних рішень. Обґрунтовано вибір методу послідовного аналізу для оцінки рівня знань у комп'ютерній системі підтримки навчальної діяльності.

Досягнуто підвищення ефективності процесу підтримки прийняття рішень під час статистичного контролю знань за рахунок удосконалення методу послідовного аналізу відповідей шляхом класифікації учасників навчального процесу на чотири групи, які відповідають рівням навченості, що забезпечило індивідуальну мінімізацію часу перевірки знань та дозволило виконати диференційовану оцінку їхнього обсягу в задачах контролю за альтернативною ознакою.

Здійснено апробацію процедури підтримки прийняття рішень на основі методу послідовного аналізу з використанням СППР «ManageEdu», в результаті чого сформовані результуючі оцінки навчальних досягнень студентів III курсу ВНЗ МТУ «Миколаївська політехніка» економічного факультету з дисципліни «Фінансовий облік» та здійснено їх класифікацію за групами, які відповідають рівням навченості. За підсумками оцінки відхилення результатів адаптивного режиму тестування від стандартного підтверджено достатню якість запропонованої класифікації.

Перспективи подальших досліджень спрямовані на розробку методики встановлення стандарту оцінювання кваліфікації (ймовірності вірно виконаних завдань для кожного класу), методика оцінки рівня помилок цього критерію, застосування запропонованої обчислювальної технології для вирішення практичних задач контролю рівня знань в системах тестування.

ПОДЯКИ

Роботу виконано в рамках спільних наукових досліджень кафедри комп'ютерної інженерії й кафедри комп'ютерних наук та прикладної математики Миколаївського національного університету ім. В. О. Сухомлинського. Результати досліджень здійснювались у рамках науко-дослідної роботи за темою: «Моделі та методи інтелектуального аналізу даних в предметно-орієнтованій інформаційній системі» (номер реєстрації 0115U001249).

СПИСОК ЛІТЕРАТУРИ

1. Кузьма К. Т. Інформаційні технології контролю та оцінки знань / К. Т. Кузьма // Труди IX Міжнародної наук.-практ. конференції студентів та молодих учених «Політ». – К. : Видво Нац. авіац. ун-ту «НАУ-друк», 2009. – С. 221.
2. Васильев В. И. Основы культуры адаптивного тестирования / В. И. Васильев, Т. Н. Тягунова. – М. : Издательство ИКАР, 2003. – 584 с.
3. Galeev I. A Learning Model in MONAP / I. Galeev, V. Ivanov, M. Akhmadullin // Human-Computer Interaction. The 6th International Conference. EWHCI'96. Moscow, 1996. – P. 320–323.
4. Автоматизация контроля обученности в процессе подготовки специалистов для систем безопасности / [А. Н. Членов, И. Г. Дровникова, Т. А. Буцынская, П. А. Орлов] // Научный информационный сборник «Проблемы безопасности и чрезвычайных ситуаций». – М. : Винити, 2009. – № 4. – С. 107–116.
5. Переверзев В. Ю. Критериально-ориентированные педагогические тесты для итоговой аттестации студентов / В. Ю. Переверзев. – М. : НМЦ СПО Минобразования РФ, 1999. – 152 с.
6. Вальд А. Последовательный анализ / А. Вальд. – М. : Физматгиз, 1960. – 328 с.
7. Левин Б. Р. Теоретические основы статистической радиотехники. В трех книгах. Книга вторая / Б. Р. Левин. – М. : Сов. радио, 1975. – 392 с.
8. Люсин Д. В. Основы разработки и применения критериально-ориентированных педагогических тестов / Д. В. Люсин. – М. : Исследовательский центр, 1993. – 51 с.
9. Angoff W. H. Scales, norms, and equivalent scores / W. H. Angoff. – Princeton, NJ: ETS, 1984. – P. 153. – URL: http://www.ets.org/Media/Research/pdf/Angoff.Scales.Norms.Equiv. Scores.pdf.
10. Kaftandjieva F. Methods for Setting Cut Scores in Criterion-referenced Achievement Tests / F. Kaftandjieva. – Cito, Arnhem: EALTA, 2010. – P. 170. – URL: http://www.ealta.eu.org/documents/resources/ FK_second_doctorate.pdf
11. А. с. 39905 України Комп'ютерна програма підтримки навчальної діяльності студентів «ManageEdu» / Кузьма К. Т., Байбуз О. Г. – № 40241; заявл. 30.06.2011; опуб. 01.09.2011.

Стаття надійшла до редакції 16.12.2017.
Після доробки 22.02.2017.

Кузьма К. Т.
Канд. техн. наук, старший преподаватель кафедры компьютерной инженерии, Николаевский национальный университет им. В. А. Сухомлинского, Николаев, Украина

ВЫЧИСЛИТЕЛЬНАЯ ТЕХНОЛОГИЯ ПРОВЕРКИ УРОВНЯ ЗНАНИЙ НА ОСНОВЕ МЕТОДА ПОСЛЕДОВАТЕЛЬНОГО АНАЛИЗА

Актуальность. Решена актуальная задача повышения эффективности процесса поддержки принятия решений при статистическом контроле знаний.

Цель работы – разработка вычислительной процедуры решения четырехальтернативной задачи классификации тестируемых по уровню обученности, что позволяет осуществлять контроль знаний дифференцированно, минимизирует объем задач, необходимый для выполнения.

Метод. Предложено вычислительную процедуру классификации тестируемых на четыре класса, которые соответствуют уровням обученности: «начальный», «средний», «достаточный», «высокий», основанную на использовании двухальтернативного критерия последовательного анализа в несколько этапов и обеспечивающую выполнение контроля знаний в процессе выполнения заданий, минимизируя таким образом время проверки знаний, что позволяет автоматизировать процесс проверки статистических гипотез в системах тестирования и обучения с целью дифференциальной оценки знаний участников учебного процесса. Для решения задачи оценки пригодности теста предложен метод, основанный на построении функции оперативной характеристики последовательного критерия, которая позволяет определить объем задач достаточный для достижения желаемого уровня качества теста за счет установления связи между ожидаемой вероятностью принятия гипотезы и случайным значением параметра вероятности появления в выборке из 1,2 ... *n* вопросов приемочного числа неверно выполненных заданий.

Результаты. Разработано программное обеспечение, которое реализует предложенную вычислительную процедуру, использованное при проведении вычислительных экспериментов тестового контроля знаний.

Выводы. Проведенные эксперименты подтвердили работоспособность предложенной процедуры и программного обеспечения, которое ее реализует, а также позволяют рекомендовать их для применения на практике для решения задач автоматизированной проверки уровня знаний.

Ключевые слова: последовательный анализ, проверка уровня знаний, классификация тестируемых по уровню знаний, проверка гипотез, критериально-ориентированный тест.

Kuzma K. T.
Ph.D., Senior Lecturer of Department of Computer Engineering, V. O. Sukhomlynsky Mykolaiv National University, Mykolaiv, Ukraine

COMPUTER TECHNOLOGIES OF VERIFICATION OF KNOWLEDGE BASED ON THE METHOD OF SEQUENTIAL ANALYSIS

Context. The actual task of increasing the effectiveness of decision support in the statistical control of knowledge has been solved.

Objective is a development of a computational procedure for solving the problem of classification the educational process participants into four groups, according to their trainability level that allows to control knowledge differently, minimizing the amount of tasks required to accomplished.

Method. The computer technology for classification the educational process participants into four groups, according to their trainability level: “initial”, “medium”, “sufficiently”, “high”, has been proposed, based on the use of sequential hypothesis testing procedure and allows to perform the controlling of the knowledge while accomplishing tasks, thus minimizing the time for testing and provides the automation of the process of verification the statistical hypotheses in a testing and learning systems with the purpose of differential assessment of knowledge the educational process participants. To solve the problem of definition the assessment standard the method based on creation the function of operational characteristics of sequential criteria is used. The function allows to establish a link between the expected probability of the hypothesis and random probability of the fact of presence in the sample with 1.2 ...*n* questions appropriate number of incorrect answers.

Results. The software implementing proposed computational procedure have been developed and used in computational experiments of knowledge testing.

Conclusions. The experiments confirmed the efficiency of the proposed procedure and software. The experiments also allow to recommend them for use in practice to solve the problems of automated assessment of knowledge.

Keywords: sequential analysis, the assessment of knowledge, classification the educational process participants according to their trainability level, testing of hypotheses, criterion-referenced test.

REFERENCES

1. Kuz’ma K.T. Informacijni tehnologiï kontrolju ta ocinki znan’, *Trudi IX Mizhnarodnoï nauk.-prakt. konferenciï studentiv ta molodih uchenih «Polit»*. Kyiv, Vid-vo Nac. aviac. un-tu «NAU-druk», 2009. – S. 221.
2. Vasil’ev V. I., Tjagunova T. N. Osnovy kul’tury adaptivnogo testirovanija. Moscow: Izdatel’stvo IKAR, 2003, 584 p.
3. Galeev I., Ivanov V., Akhmadullin M. A Learning Model in MONAP, *Human-Computer Interaction. The 6th International Conference, EWHCI’96*. Moscow, 1996, pp. 320–323.
4. Chlenov A.N., Drovnikova I.G., Bucynskaja T.A., Orlov P.A. Avtomatizacija kontrolja obuchennosti v processe podgotovki specialistov dlja sistem bezopasnosti, *Nauchnyj informacionnyj sbornik «Problemy bezopasnosti i chrezvychajnyh situacij»*. Moscow, Viniti, 2009. No. 4, pp. 107–116.
5. Pereverzev V. Ju. Kriterial’no-orientirovannye pedagogicheskie testy dlja itogovoj attestacii studentov. Moscow, NMC SPO Minobrazovanija RF, 1999, 152 p.
6. Val’d A. Posledovatel’nyj analiz. Moscow, Fizmatgiz, 1960, 328 p.
7. Levin B. R. Teoreticheskie osnovy statisticheskoj radiotehniki. V treh knigah. Kniga vtoraja. Moscow, Sov. radio, 1975, 392 p.
8. Ljusin D.V. Osnovy razrabotki i primenenija kriterial’no-orientirovannyh pedagogicheskih testov. Moscow, Issledovatel’skij centr, 1993, 51 p.
9. Angoff W. H. Scales, norms, and equivalent scores. Princeton, NJ, ETS, 1984, P. 153. URL: http://www.ets.org/Media/Research/pdf/Angoff.Scales.Norms.Equiv. Scores.pdf.
10. Kaftandjieva F. Methods for Setting Cut Scores in Criterion-referenced Achievement Tests. Cito, Arnhem: EALTA, 2010, P. 170. URL: http://www.ealta.eu.org/documents/resources/FK_second_doctorate.pdf
11. Kuz’ma K. T., Bajbuz O. G. A. s. 39905 Ukraïni Komp’juterna programa pidtrimki navchal’noï dijal’nosti studentiv «ManageEdu» / № 40241; zajavl. 30.06.2011; opub. 01.09.2011.

УДК 004.056

Фауре Э. В.

Канд. техн. наук, доцент, докторант, доцент кафедры информационной безопасности и компьютерной инженерии Черкасского государственного технологического университета, Черкассы, Украина

ФАКТОРИАЛЬНОЕ КОДИРОВАНИЕ С ИСПРАВЛЕНИЕМ ОШИБОК

Актуальность. Факториальное кодирование данных позволяет совмещать операции крипто- и имитозащиты, а также помехоустойчивого кодирования, что приводит к уменьшению вносимой передатчиком избыточности, повышению быстродействия и увеличению эффективной пропускной способности. Вместе с тем описанные методы факториального кодирования не позволяют исправлять ошибки, что ограничивает область их использования.

Целью данной работы является разработка метода факториального кодирования с восстановлением данных по перестановке, обеспечивающего комплексное решение задач криптографической защиты и помехоустойчивого кодирования и позволяющего совместить функции исправления и обнаружения ошибок канала связи.

Метод. Основная идея предложенного метода кодирования состоит в увеличении расстояния между разрешенными кодовыми словами, представляющими собой перестановки, вычисленные по всем информационным битам блока данных и представленные в двоичном виде. Исследованы методы увеличения расстояния на основе метрик Эвклида и Хэмминга. Для каждого из этих методов определены основные свойства факториального кода с исправлением ошибок, в том числе выполнена оценка достоверности передачи при независимости и биномиальном распределении возникающих в канале связи ошибок, разработаны структурные схемы приемника. Правила декодирования, реализованные в приемнике, основываются на критерии максимального правдоподобия и предусматривают как прямое исправление ошибок, так и их обнаружение с последующим исправлением путем переспроса поврежденного блока.

Результаты. Реализованы факториальные коды с исправлением ошибок, использующие метрики Эвклида и Хэмминга. Для этих кодов выполнен сравнительный анализ вероятности необнаруженной ошибки, остаточной вероятности ошибочного приема, энергетического выигрыша и относительной скорости передачи. Показано, что характеристики кода не являются инвариантными по отношению к множеству разрешенных кодовых слов, а из рассмотренных в работе кодов более эффективными являются коды, использующие метрику Хэмминга.

Выводы. Получил дальнейшее развитие метод факториального кодирования с восстановлением данных по перестановке, который за счет совмещения функций исправления и обнаружения ошибок позволяет повысить динамическую составляющую потери скорости и, как следствие, относительную скорость передачи, по сравнению с обнаруживающим ошибки факториальным кодированием за счет снижения его помехоустойчивости. Проведенные эксперименты подтвердили эффективность факториальных кодов с исправлением ошибок.

Ключевые слова: избыточность, факториальный код, перестановка, помехоустойчивое кодирование, исправление ошибок, обнаружение ошибок, достоверность передачи, относительная скорость передачи.

НОМЕНКЛАТУРА

FCDR – Factorial Code with Data Recovery by permutation;

FCDRec – FCDR with error correction;

РОС – решающая обратная связь;

СКК – сигнально-кодовая конструкция;

ФКВД – факториальный код с восстановлением данных по перестановке;

ФКВДио – факториальный код с восстановлением данных и исправлением ошибок;

ΔP – энергетический выигрыш при применении помехоустойчивого кода;

α – показатель избыточности кода (по мощности);

ν_2 – динамическая составляющая потери скорости;

$A(x)$ – представленное в двоичном виде информационное слово;

D_i – эвклидово расстояние от нуля до сигнальной точки i;

$D_{i,j}$ – эвклидово расстояние между сигнальными точками i и j;

$D_{\min}$ – минимальное эвклидово расстояние между сигнальными точками;

$d_{i,j}$ – расстояние Хэмминга между сигнальными точками i и j;

$d_{\min}$ – минимальное расстояние Хэмминга между кодовыми словами;

$f_{per}^{EC}(i,t)$ – количество ошибок веса t, исправляемых для i-ой сигнальной точки;

$f_{per}^{ud}(i,t)$ – количество ошибок веса t, приводящих к ошибочному декодированию i-ого сигнального вектора

k – число двоичных символов в информационном блоке данных;

l_r – число бит для кодирования одного символа перестановки;

M – порядок перестановки;

P_{det} – вероятность обнаруженной ошибки;

P_{EC} – вероятность того, что ошибка будет исправлена, а кодовая комбинация принята верно;

P_{res} – остаточная вероятность ошибочного приема;

P_{ud} – вероятность необнаруженной ошибки;

$P_{ud}(FCDR(ec), p_0)$ – вероятность не обнаруженной ФКВД (или ФКВДио) ошибки;

$P_w(i)$ – вероятность применения источником i-ого слова;

p_0 – переходная вероятность симметричного двоичного постоянного канала;

p_{0eq} – эквивалентная вероятность битовой ошибки;

Q – вероятность приема блока данных без ошибок

DOI 10.15588/1607-3274-2017-3-15

$R_{FCDR}(x)$ – представленное в двоичном виде кодовое слово ФКВД;

r – длина кодового слова;

r_i – расстояние Хэмминга между принятым вектором и i-ым сигнальным вектором;

S_F – синдром перестановки.

ВВЕДЕНИЕ

Проблемы повышения эффективности систем передачи данных, включая повышение достоверности передачи и пропускной способности, всегда привлекали внимание специалистов информационных технологий и телекоммуникаций. В данном контексте перспективными являются методы факториального кодирования информации [1–6], позволяющие совместить операции помехоустойчивого кодирования, крипто- и имитозащиты и тем самым уменьшить вносимую передатчиком избыточность, повысить быстродействие и увеличить эффективную пропускную способность. Вместе с тем возможности факториального кодирования, изложенные в [1–6], далеко не исчерпаны, что и определяет круг решаемых в данной работе задач.

В работе [4] предложен, а в работе [5] получил дальнейшее развитие метод факториального кодирования с восстановлением данных по перестановке (ФКВД, FCDR). Данный метод направлен на комплексную защиту информации от несанкционированного чтения и ошибок, возникающих в канале связи. При этом ФКВД решает задачу обнаружения ошибок, а их исправление достигается повторной передачей искаженного помехой кодового слова. Вместе с тем при определенных обстоятельствах относительная скорость передачи информации в системах с решающей обратной может оказаться чрезмерно малой. Для ее повышения (а также в системах реального масштаба времени) целесообразно использовать коды с исправлением ошибок. Кроме того, как сказано в [7], комбинирование процедур обнаружения и исправления ошибок является зачастую более эффективной, чем либо только исправление ошибок, либо только обнаружение ошибок с повторной передачей. Поэтому актуальной задачей является задача совмещения процедур криптографического преобразования информации, а также обнаружения и исправления ошибок.

Целью данной работы является разработка метода факториального кодирования информации, который реализует функцию защиты информации от несанкционированного доступа, а также функцию помехоустойчивого кодирования, сочетающего обнаружение и исправление ошибок, возникающих в канале связи.

1 ПОСТАНОВКА ЗАДАЧИ

Пусть информация от источника поступает на вход кодера блоками из k бит. Тогда мощность множества информационных слов составляет 2^k. Обозначим через $A(x)$ представленное в двоичном виде информационное слово (вектор). ФКВД реализует биективное преобразование множества информационных слов $A(x)$ в разрешенное множество из 2^k перестановок $R_{FCDR}(x)$ порядка M $\left(M! \geq 2^k\right)$.

Задачей синтеза метода факториального кодирования информации, сочетающего защиту информации от несанкционированного чтения, а также обнаружение и исправление ошибок заключается в определении и анализе структуры множества из 2^k векторов $R_{FCDR}(x)$, позволяющей выделить для каждого из них область возможных значений $R^{\wedge}_{FCDR}(x)$ на входе приемника, расстояние до которых в заданной метрике не превышает заданного значения, а также область значений $R^{\wedge}_{FCDR}(x)$, находящихся на одинаковом расстоянии до двух или более векторов $R_{FCDR}(x)$ из 2^k возможных.

2 ОБЗОР ЛИТЕРАТУРЫ

Согласно [4], перестановка $R_{FCDR}(x)$ представляет собой последовательность закодированных равномерным двоичным кодом чисел $\{0;1;\ldots;M-1\}$, очередность следования которых определяется информационной последовательностью и алгоритмом кодирования. Если порядок формирования перестановки по информационному слову источника держится в секрете, ФКВД, помимо обнаружения ошибок в канале связи, обеспечивает защиту данных от несанкционированного чтения. Кроме того, такой код является самосинхронизирующимся и не требует разделителя кодовых слов.

Как показано в [4], приемник содержит блок проверки корректности принятой из канала кодовой комбинации и декодер ФКВД. Проверка корректности сводится к проверке того факта, что в принятой кодовой комбинации каждый символ множества $\{0;1;\ldots;M-1\}$ применяется ровно по одному разу. В случае, если принятая последовательность является некорректной, она не допускается к декодированию, а на передающую станцию по обратному каналу связи передается запрос повторной передачи блока.

Корректная последовательность подлежит декодированию – обратному преобразованию $f^{-1}_{FCDR}: R_{FCDR}(x) \to A(x)$. Согласно [4], поскольку $M! > 2^k$ при $k > 1$, множество перестановок на входе декодера состоит из двух подмножеств – разрешенного и запрещенного. К разрешенному подмножеству относятся 2^k перестановок (в простейшем случае их синдромы S_F соответствуют целым числам $\left[0; 2^k - 1\right]$ числовой оси), а к запрещенному – подмножество из $\left(M! - 2^k\right)$ остальных перестановок (в простейшем случае их синдромы S_F соответствуют целым числам $\left[2^k; M! - 1\right]$ числовой оси). Прием любой перестановки из неразрешенной части множества также инициирует команду переспроса.

В работе [5] для ФКВД введен показатель избыточности (по мощности) α:

$$\alpha = M! / 2^k . \qquad (1)$$

В [5] также показано, что при $k>1$ справедливо $M!>2^k$ и, соответственно, $\alpha>1$, что приводит к избыточности кода. При этом введение дополнительных проверочных бит перед преобразованием информационного вектора в перестановку позволяет повысить обнаруживающую способность ФКВД. С другой стороны, избыточность ФКВД обеспечивает возможность увеличения расстояния между перестановками – носителями информации и создает предпосылки для создания факториального кода с исправлением ошибок. Такой код будем называть факториальным кодом с восстановлением данных и исправлением ошибок – ФКВДио (FCDRec – FCDR with error correction).

3 МАТЕРИАЛЫ И МЕТОДЫ

Введем следующие определения.

Определение 1. Сигнальными векторами называются представленные в двоичном виде перестановки разрешенного множества.

Множество сигнальных векторов образует сигнально-кодовую конструкцию (СКК).

Определение 2. Сигнальными точками называются точки на числовой оси $[0;M!-1]$, которые соответствуют сигнальным векторам кода.

Множество сигнальных точек кода образует его сигнальное созвездие.

Рассмотрим два способа формирования СКК для ФКВДио:

1) СКК, основанная на минимальном расстоянии Эвклида между сигнальными точками. Такие СКК будем называть СКК первого типа и обозначать через СКК-1;

2) СКК, основанная на минимальном расстоянии Хэмминга между сигнальными векторами. Такие СКК будем называть СКК второго типа и обозначать через СКК-2.

Рассмотрим ФКВДио с СКК-1.

Минимальное расстояние между сигнальными точками на оси $[0;M!-1]$

$$D_{\min}\le\left[(M!-1)/(2^k-1)\right]. \qquad (2)$$

В простейшем случае 2^k сигнальных точек располагаются на числовой оси $[0;2^k-1]$ с шагом $D_{\min}=1$. Такой код не предназначен для исправления ошибок. Он может быть применен для обнаружения ошибок, причем только тех, которые приводят к преобразованию перестановки в «не перестановку» или в перестановку из запрещенного множества.

Напомним, что для ФКВД $k\le[\log_2 M!]$. Выполним оценку $D_{\min}$ при $k=[\log_2 M!]$, для которого достигается максимальная скорость кода. Поскольку $\log_2 M!-1<[\log_2 M!]\le\log_2 M!$, имеет место $M!/2<2^k\le M!$. Тогда $1\le(M!-1)/(2^k-1)<2+2/(M!-2)$. Поэтому при $k=[\log_2 M!]$ минимальное расстояние между сигнальными точками $D_{\min}\le 2$. Такой ФКВД не способен исправлять все ошибки, приводящие даже к минимальному смещению сигнальных векторов по числовой оси, и поэтому его целесообразно применять для обнаружения ошибок. При этом, как и для $D_{\min}=1$, обнаруживаются только те ошибки, которые приводят к преобразованию переданной перестановки в «не перестановку» или в перестановку из запрещенного множества.

Исправление ошибок возможно при $D_{\min}\ge 3$. Для увеличения $D_{\min}$ необходимо увеличивать показатель α. Очевидно, что $\alpha=M!/2^k$ монотонно возрастает по *M* и убывает по *k*. Поэтому увеличение может быть достигнуто как увеличением *M*, так и уменьшением *k*. При этом, как показано в [5], уменьшение длины информационного вектора на Δk бит при фиксированном *M* приводит к увеличению показателя α в $2^{\Delta k}$ раз.

Графически расположение сигнальных точек на числовой оси представлено на рис. 1. При этом расстояние от нуля до сигнальной точки *i* будем обозначать через D_i, а между сигнальными точками *i* и *j* – через $D_{i,j}=D_j-D_i$.

Положение сигнальных точек на числовой оси определяется СКК. В простейшем случае сигнальные точки располагаются равномерно с шагом $D_{\min}$, при этом $D_{i,i+1}=D_{\min}$ для $i\in[0;2^k-2]$. В более общем случае $D_{i,i+1}\ne\text{const}$, а $D_{i,j}\ge D_{\min}$, $i,j\in[0;2^k-1]$, $i\ne j$.

При передаче сигнального вектора действующая в канале связи помеха может сместить сигнальную точку передатчика в любую другую точку отрезка $[0;M!-1]$, которая может быть как сигнальной, так и не сигнальной, а принятая перестановка может относиться как к разрешенному, так и к запрещенному множеству.

Приемник принимает решение о переданном сигнальном векторе на основании критерия максимального правдоподобия путем нахождения сигнальной точки, ближайшей (в метрике Эвклида) к точке, соответствующей принятому вектору. Для этого декодер вычисляет расстояния от соответствующей принятому вектору точке числовой оси до соседних сигнальных точек. При равенстве этих расстояний формируется сигнал переспроса.

Таким образом, если помеха сместила сформированный передатчиком *i*-ый вектор не более чем на $-[(D_{i-1,i}-1)/2]$ и $+[(D_{i,i+1}-1)/2]$ точек, эта ошибка исправляется, а принятый вектор корректируется приемником в перестановку, соответствующую *i*-ой сигнальной точке. Если смещение равняется $-[D_{i-1,i}/2]$ (или $+[D_{i,i+1}/2]$) и при этом $D_{i-1,i}/2\in\mathbb{Z}$ $(D_{i,i+1}/2\in\mathbb{Z})$, ошибка обнаруживается кодом и исправляется пере-

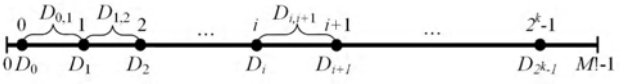


Рисунок 1 – Расположение сигнальных точек на числовой оси

спросом. Если же смещение превышает $-\left[D_{i-1,i}/2\right]$ (или $+\left[D_{i,i+1}/2\right]$), такие ошибки код исправить не может. В этом случае, если расстояния от соответствующей принятому вектору точки до соседних сигнальных точек одинаковы, ошибка обнаруживается, в противном случае возникает ошибка декодирования и, как следствие, не обнаруженная кодом ошибка.

Если принятый вектор соответствует точке из диапазона $\left[D_0-\left[\left(D_{\min}-1\right)/2\right];D_0-1\right]$, приемник корректирует ее в нулевую сигнальную точку, если же точке из диапазона $\left[D_{2^k-1}+1;D_{2^k-1}+\left[\left(D_{\min}-1\right)/2\right]\right]$ – в $\left(2^k-1\right)$ сигнальную точку, остальные точки диапазонов $\left[0;D_0-1\right]$ или $\left[D_{2^k-1}+1;M!-1\right]$ являются запрещенными.

Таким образом, все ошибки, приводящие к смещению сигнальной точки на расстояние $D\le\left[\left(D_{\min}-1\right)/2\right]$, исправляются кодом.

Положим $D_{i,i+1}=D_{\min}$ для $\forall i\in\left[0;2^k-2\right]$, $D_0=\left[\left(D_{\min}-1\right)/2\right]$, а $M!-1-D_{2^k-1}\ge\left[\left(D_{\min}-1\right)/2\right]$. В этом случае имеет место оценка

$$\left(2^k-1\right)D_{\min}+2\left[\left(D_{\min}-1\right)/2\right]+1\le M!. \qquad (3)$$

При заданных k и M $D_{\min}\le\max\left(D\right):\left(2^k-1\right)D+2\left[\left(D-1\right)/2\right]+1\le M!$.

Например, если $k=40$, а $M=16$, то $D_{\min}\le 19$. Поэтому выбор параметров k и M однозначно определяют максимальную исправляющую способность кода. Выражение (3) также может служить для выбора k или M при других известных параметрах. Например, если $k=16$, $D_{\min}=3$, то $M\ge 9$; если $M=8$, $D_{\min}=6$, то $k\le 12$. Кроме того, выражение (3) показывает, что все точки, лежащие правее пороговой точки $\left(2^k-1\right)D_{\min}+2\left[\left(D_{\min}-1\right)/2\right]+1$, относятся к запрещенной части числового множества. Поэтому все принятые кодовые комбинации после проверки корректности проходят сравнение с пороговым значением. Если соответствующая кодовой комбинации точка расположена выше пороговой точки, производится переспрос блока данных, в противном случае выполняется поиск ближайшей сигнальной точки и отождествление с ней принятой кодовой комбинации.

Структурная схема приемника ФКВДио представлена на рис. 2, где введены следующие обозначения: БПК – блок проверки корректности принятой комбинации; БИИ – блок извлечения информации из перестановки; БОМ – блок оценки принадлежности принятой перестановки к разрешенному множеству; БО – блок отождествления принятой перестановки с ближайшим разрешенным вектором данных.

Определим вероятностные характеристики ФКВДио с СКК-1.

Примем, что канал связи – симметричный двоичный постоянный с переходной вероятностью p_0, а битовые ошибки возникают в нем независимо. Тогда вероятность не обнаруженной ФКВД или ФКВДио ошибки

$$P_{ud}\left(FCDR(ec),p_0\right)=\sum_{i=0}^{2^k-1}\left(P_w\left(i\right)\cdot\sum_{t=1}^{r}f_{per}^{ud}\left(i,t\right)p_0^t q_0^{r-t}\right). \qquad (4)$$

Доля ошибок, приводящих к ошибочному декодированию: $\left(2^k-1\right)/M!$ для ФКВД и $\left(2^k-1\right)\left(2\left[\left(D_{\min}-1\right)/2\right]+1\right)/M!$ для ФКВДио. Поскольку множество ошибок, приводящих к ошибочному декодированию ФКВДио, содержит множество ошибок, приводящих к ошибочному декодированию ФКВД, $P_{ud}\left(FCDRec,p_0\right)>P_{ud}\left(FCDR,p_0\right)$ при $D_{\min}\ge 3$.

Учтем, что для простейшей системы с РОС динамическая составляющая потери скорости вследствие переспросов $\nu_2=Q+P_{ud}$ [8]. В случае использования ФКВДио

$$Q+P_{EC}+P_{det}+P_{ud}=1. \qquad (5)$$

Тогда динамическая составляющая потери скорости для ФКВДио:

$$\nu_2=1-P_{det}=Q+P_{EC}+P_{ud}. \qquad (6)$$

Вероятность исправления ошибок для ФКВДио:

$$P_{EC}\left(FCDRec,p_0\right)=\sum_{i=0}^{2^k-1}\left(P_w\left(i\right)\cdot\sum_{t=1}^{r}f_{per}^{EC}\left(i,t\right)p_0^t q_0^{r-t}\right). \qquad (7)$$

Поскольку $\nu_2=Q+P_{ud}$ для ФКВД меньше значения ν_2 по (6) для ФКВДио, при одинаковых параметрах и $D_{\min}\ge 3$ ФКВДио обеспечивает большую относительную скорость передачи по сравнению с ФКВД, однако проигрывает в помехоустойчивости.

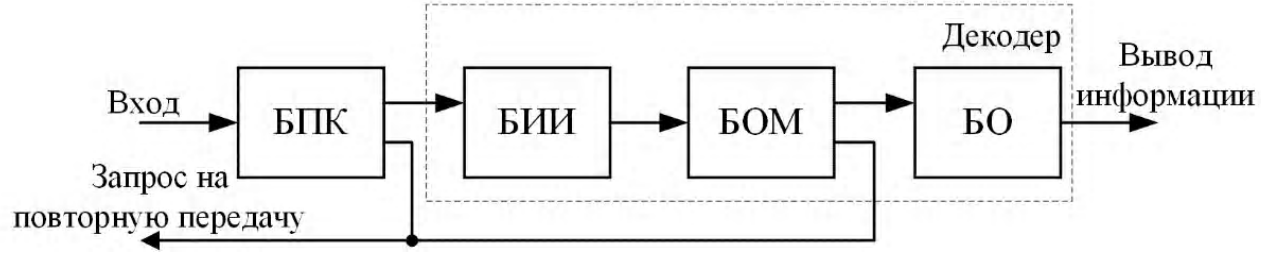


Рисунок 2 – Структурная схема приемника ФКВДио

Определим энергетический выигрыш ΔP при применении ФКВДио для некогерентного приемника, характеризующегося вероятностью битовой ошибки $p = 0,5 \cdot e^{-0,5h^2}$ [9], где h^2 – соотношение сигнал/шум. В этом случае

$$\Delta P = 10\lg\left(\ln h_{eq}^2 / \ln h_0^2\right) = 10\lg\left(\ln\left(2p_{0eq}\right)/\ln\left(2p_0\right)\right), \quad (8)$$

где эквивалентная вероятность битовой ошибки p_{0eq} [8] равна

$$p_{0eq} \approx P_{ud} / \left(k\left(1 - P_{det}\right)\right). \quad (9)$$

Учтем, что из (5) $1 - P_{det} = Q + P_{EC} + P_{ud}$. Тогда выражение (9) принимает вид:

$$p_{0eq} \approx P_{ud} / \left(k\left(Q + P_{EC} + P_{ud}\right)\right). \quad (10)$$

Остаточная вероятность ошибочного приема [8] для ФКВДио равна

$$P_{res} = P_{ud} / \left(1 - P_{det}\right) = P_{ud} / \left(Q + P_{EC} + P_{ud}\right). \quad (11)$$

Рассмотрим ФКВДио с СКК-2.

Определенное для СКК-1 расстояние Эвклида $D_{i,j}$ между сигнальными точками *i* и *j* в общем случае не равняется расстоянию Хэмминга между кодовыми словами, соответствующими этим сигнальным точкам. Вместе с тем из теории корректирующих кодов [7] известно, что для исправления ошибки в двоичных разрядах кратности *t* минимальное расстояние Хэмминга между кодовыми словами $d_{\min} \geq 2t + 1$.

Расстояние Хэмминга между сигнальными векторами *i* и *j* будем обозначать через $d_{i,j}$. СКК-2 для ФКВДио предусматривает выполнение условия $d_{i,j} \geq d_{\min}$, $i, j \in \left[0; 2^k - 1\right]$, $i \neq j$.

Определение связи между *M*, *k* и $d_{\min}$ является актуальной задачей, однако выходит за рамки данной работы. В простейшем случае для ФКВД сигнальные вектора соответствуют сигнальным точкам с шагом $D_{i,i+1} = D_{\min} = 1$ и $D_0 = 1$. Тогда $d_{\min} = 2$, а ФКВД только обнаруживает ошибки, приводящие к преобразованию переданной перестановки в «не перестановку» или в перестановку из запрещенного множества. Очевидно, что для обеспечения возможности исправления ошибок необходимо увеличивать $d_{\min}$. Для этого необходимо увеличивать показатель α. При этом ошибки исправляются при $d_{\min} \geq 3$.

При передаче сигнального вектора по каналу связи на него воздействует помеха. Модифицированный помехой вектор поступает на вход приемника ФКВДио с СКК-2, структура которого показана на рис. 3. Приемник содержит блок исправления и обнаружения ошибок БИОО и блок извлечения информации из перестановки БИИ.

БИОО реализует следующие функции: определяет расстояния Хэмминга r_i между принятым вектором и всеми сигнальными векторами, $i \in \left[0; 2^k - 1\right]$; находит минимальное расстояние $r_{\min} = \min\left\{r_i\right\}$; если существует единственное $i \in \left[0; 2^k - 1\right]$: $r_i = r_{\min}$, принятая комбинация отождествляется с *i*-ым сигнальным вектором; если существует как минимум два значения $i, j \in \left[0; 2^k - 1\right]$: $r_i = r_j = r_{\min}$, формируется сигнал переспроса. Таким образом, правила декодирования основываются на критерии максимального правдоподобия.

В БИИ производится преобразование перестановки в *k*-битную последовательность.

Учтем, что $d_{i,j} \vdots 2$ и, следовательно, $d_{\min} \vdots 2$. Поэтому при передаче *i*-го сигнального вектора ФКВДио с СКК-2 справедливы следующие утверждения:

1) ошибка с весом $t \leq \left[\left(d_{\min} - 1\right)/2\right] = \left(d_{\min} - 2\right)/2 = d_{\min}/2 - 1$ исправляется, а принятый вектор корректируется приемником в переданный сигнальный вектор;

2) ошибка с весом $t = d_{\min}/2$ может быть как исправлена (если $r_{\min}$ соответствует расстоянию только до одного *i*-го сигнального вектора), так и обнаружена и исправлена переспросом (если $r_{\min}$ соответствует расстоянию до двух и более сигнальных векторов);

3) если $t > d_{\min}/2$, ошибка либо исправляется (если $r_{\min}$ соответствует расстоянию только до одного *i*-го сигнального вектора), либо обнаруживается (если соответствует расстоянию до двух и более сигнальных векторов), либо не обнаруживается (если соответствует расстоянию только до одного сигнального вектора, отличного от *i*-го).

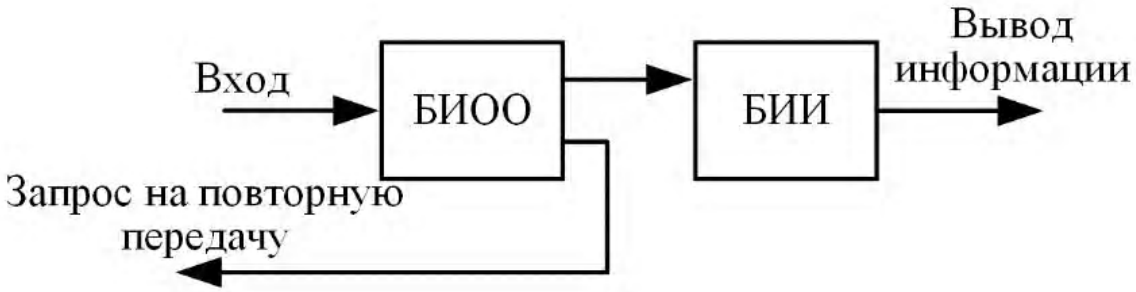


Рисунок 3 – Структурная схема приемника ФКВДио с СКК-2

Таким образом, ФКВДио с СКК-2 позволяет комбинировать исправление наиболее частых сочетаний ошибок и обнаружение с последующей повторной передачей для более редких сочетаний ошибок. Вероятностные характеристики определяются по тем же формулам, что и ФКВДио с СКК-1: (4), (6), (7), (8), (11).

Актуальным, однако выходящим за рамки данной работы, является вопрос о максимальном количестве $N_{sv}\left(d_{\min}, M\right)$ сигнальных векторов, обеспечивающих заданное $d_{\min}$ при известном M (данный вопрос тесно связан с теорией решеток и задачей наилучших упаковкой шаров в пространствах различных размерностей [10]).

4 ЭКСПЕРИМЕНТЫ

Примем $k=3$, а $M=4$. В соответствии с (3) для СКК-1 $D_{\min} \le 3$. Тогда сигнальными точками для ФКВДио с СКК-1 являются точки 1, 4, 7, 10, 13, 16, 19, 22. СКК-1 для базовой перестановки $\pi(0)=\{0;1;2;3\}$ представлена в табл. 1.

Ошибка не обнаруживается кодом, если кодовое слово, соответствующее *i*-ой сигнальной точке СКК-1, будет преобразовано помехой в комбинацию, соответствующую любой точке числовой оси, не принадлежащей диапазону $\left[D_i-1; D_i+1\right]$.

Примем, что все слова применяются источником с одинаковой вероятностью $P_w(i)=P_w=1/2^k$. Для $k=3$, $M=4$: $P_w=1/8$, $l_r=2$, $r=8$. Учтем, что $f_{per}^{ud}(i,t)=0$ при $t<2$. Тогда $P_{ud}\left(FCDRec, p_0\right)=1/8\sum_{i=0}^{7}\sum_{t=1}^{4} f_{per}^{ud}(i,2t)\, p_0^{2t} q_0^{8-2t}$. Значения $f_{per}^{ud}(i,2t)$ приведены в табл. 2.

С учетом четности ошибок, преобразующих перестановку в перестановку, для СКК-1 $P_{EC}\left(FCDRec, p_0\right)=\sum_{i=0}^{2^k-1}\left(P_w(i)\cdot\sum_{t=1}^{[r/2]} f_{per}^{EC}(i,2t)\, p_0^{2t} q_0^{r-2t}\right)$.

Значения $f_{per}^{EC}(i,2t)$ приведены в табл. 3.

Значения $f_{per}^{ud}(i,2t)$ для ФКВД с СКК-1 приведены в табл. 4.

В табл. 1 представлена СКК-2 с $d_{\min}=4$. Экспериментально установлено, что ФКВДио с такой СКК исправляет только любые ошибки с $t=1$, а ошибка не обнаруживается тогда и только тогда, когда кодовое слово, соответствующее *i*-ому сигнальному вектору, преобразовывается помехой в комбинацию, для которой $r_j \le 1$ для $j \in \left[0; 2^k-1\right]$, $j \ne i$. Поэтому $f_{per}^{EC}(i,t)=8$ при $t=1$ и $f_{per}^{EC}(i,t)=0$ при $t \ne 1$.

Значения $f_{per}^{ud}(i,t)$ для СКК-2 не зависят от сигнальной точки *i* и равны: $f_{per}^{ud}(i,3)=24$, $f_{per}^{ud}(i,4)=6$, $f_{per}^{ud}(i,5)=24$, $f_{per}^{ud}(i,6)=0$, $f_{per}^{ud}(i,7)=8$, $f_{per}^{ud}(i,8)=1$, $f_{per}^{ud}(i,t)=0$ при $t \le 2$.

Для обнаруживающего ошибки ФКВД с СКК-2 значения $f_{per}^{ud}(i,t)$ не зависят от сигнальной точки *i* и равны: $f_{per}^{ud}(i,4)=6$, $f_{per}^{ud}(i,8)=1$, $f_{per}^{ud}(i,t)=0$ при других *t*.

5 РЕЗУЛЬТАТЫ

На рис. 4 показаны графики зависимостей вероятностей необнаруженной ошибки от вероятности битовой ошибки p_0 для рассмотренных кодов: ФКВДио с СКК-1 (FCDRec-1) и СКК-2 (FCDRec-2), а также ФКВД с СКК-1 (FCDR-1) и СКК-2 (FCDR-2).

На рис. 5 для этих кодов показаны графики зависимостей величины $1-\nu_2$ от p_0.

Таблица 2 – Значения $f_{per}^{ud}(i,2t)$ для ФКВДио с СКК-1

t	Сигнальная точка							
	0	1	2	3	4	5	6	7
1	3	4	3	3	3	3	4	3
2	13	12	13	13	13	13	12	13
3	4	4	4	4	4	4	4	4
4	1	1	1	1	1	1	1	1

Таблица 3 – Значения $f_{per}^{EC}(i,2t)$ для ФКВДио с СКК-1

t	Сигнальная точка							
	0	1	2	3	4	5	6	7
1	1	0	1	1	1	1	0	1
2	1	2	1	1	1	1	2	1

Таблица 4 – Значения $f_{per}^{ud}(i,2t)$ для ФКВД с СКК-1

t	Сигнальная точка							
	0	1	2	3	4	5	6	7
1	2	2	1	1	1	1	2	2
2	2	2	4	4	4	4	2	2
3	2	2	1	1	1	1	2	2
4	1	1	1	1	1	1	1	1

Таблица 1 – СКК-1 и СКК-2 для ФКВДио при $k=3$, $M=4$

СКК-1		СКК-2	
Сигнальные точки	Сигнальные вектора	Сигнальные вектора	Сигнальные точки
1	00 01 11 10	00 01 10 11	0
4	00 11 01 10	01 00 11 10	7
7	01 00 11 10	10 11 00 01	16
10	01 11 00 10	11 10 01 00	23
13	10 00 11 01	11 01 10 00	21
16	10 11 00 01	01 11 00 10	10
19	11 00 10 01	10 00 11 01	13
22	11 10 00 01	00 10 01 11	2

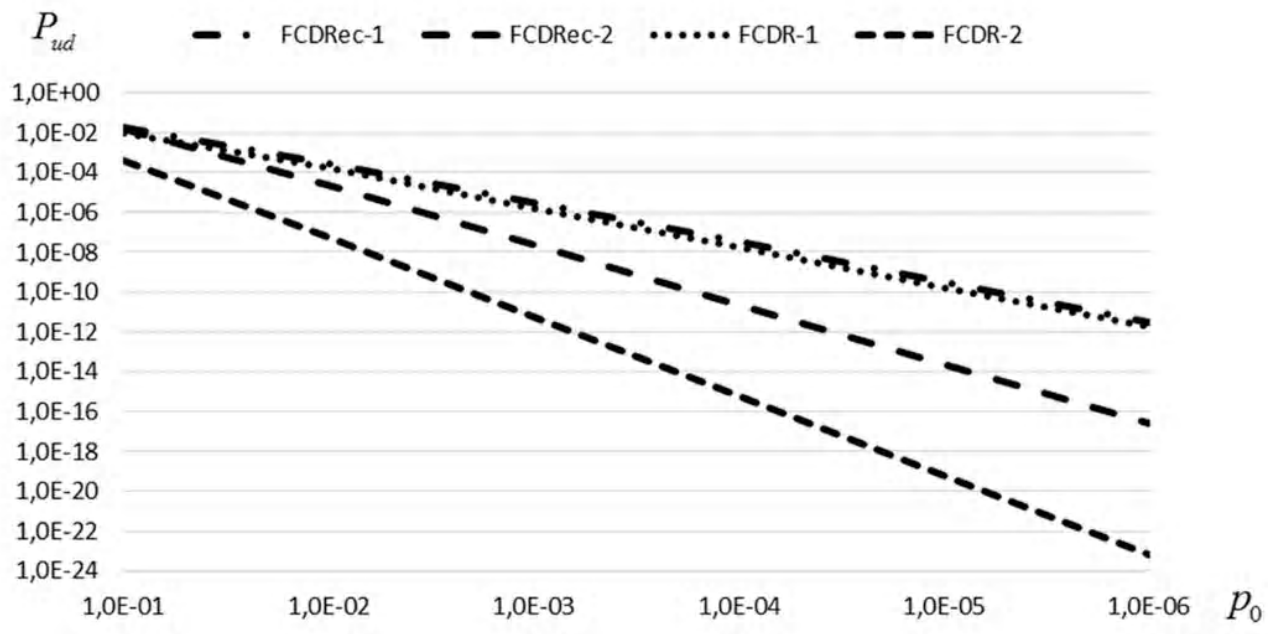


Рисунок 4 – Графики зависимостей вероятностей необнаруженной ошибки от p_0

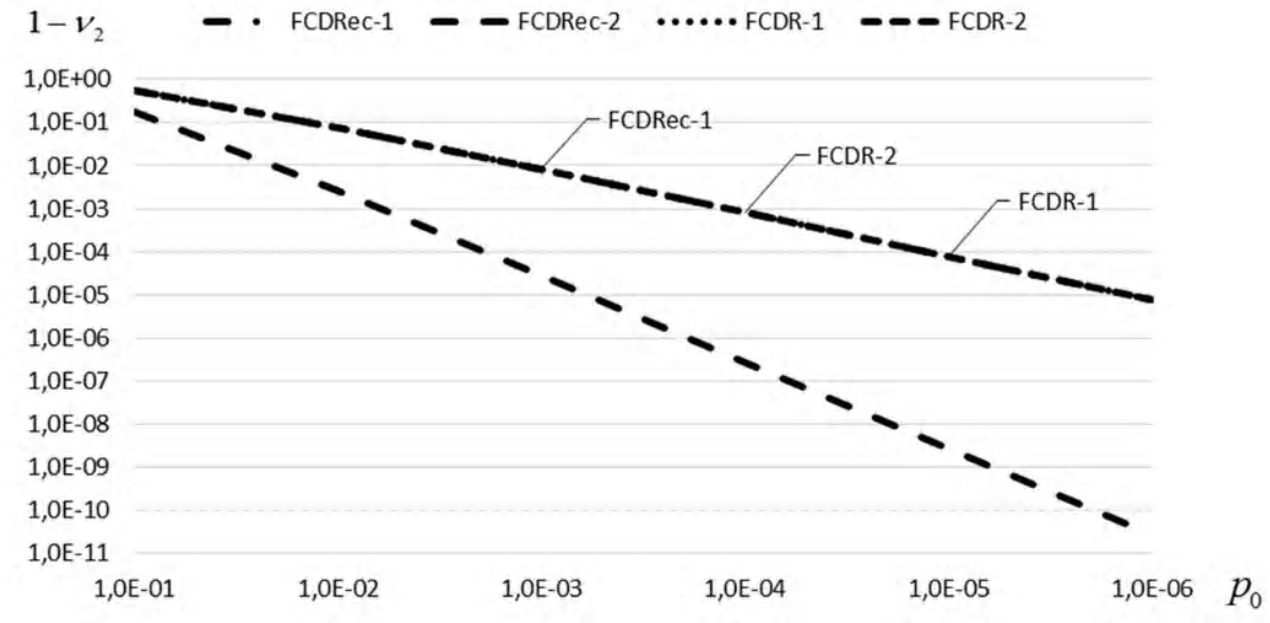


Рисунок 5 – Графики зависимостей величины $1-\nu_2$ от p_0

6 ОБСУЖДЕНИЕ

Из представленных графиков следует, что наименьшую достоверность передачи из рассмотренных кодов обеспечивает ФКВДио с СКК-1, для которого вероятность необнаруженной ошибки практически в два раза больше по сравнению с ФКВД с СКК-1. При этом динамическая составляющая потери скорости для ФКВДио и ФКВД с СКК-1 различаются незначительно (менее 3% при $p_0 = 0{,}1$). Поэтому в данном сравнении ФКВД с СКК-1 является более предпочтительным. Вместе с тем из этого пока не следует вывод в общем случае о меньшей эффективности ФКВДио по сравнению с ФКВД для СКК-1.

Наименьшую вероятность необнаруженной ошибки и наибольший энергетический выигрыш имеет ФКВД с СКК-2. Отношение $P_{ud}\left(FCDR-2, p_0\right)/P_{ud}\left(FCDR-1, p_0\right)$ равно $2{,}4\cdot 10^1$ при $p_0 = 0{,}1$ (разница в энергетическом выигрыше $\Delta P = 2{,}42$ дБ) и $2{,}9\cdot 10^5$ при $p_0 = 0{,}001$ ($\Delta P = 2{,}84$ дБ), указывая на большую эффективность СКК-2 для ФКВД.

Использование СКК-2 для ФКВДио увеличивает по сравнению с ФКВД вероятность необнаруженной ошибки: $P_{ud}\left(FCDRec-2, p_0\right)/P_{ud}\left(FCDR-2, p_0\right)$ равно $3{,}7\cdot 10^1$

при $p_0 = 0,1$ ($\Delta P = 2,23$ дБ) и $4 \cdot 10^3$ при $p_0 = 0,001$ ($\Delta P = 1,65$ дБ). Вместе с тем $\nu_2\left(FCDR-2, p_0\right) \approx \nu_2\left(FCDR\left(ec\right)-1, p_0\right)$, в то время, как $\nu_2\left(FCDRec-2, p_0\right)$ может значительно превосходить $\nu_2\left(FCDR-2, p_0\right)$: их разность при $p_0 = 0,1$ равна $0,828-0,431 \approx 0,4$ (более 92%) и увеличивается с увеличением p_0. Применение СКК-2 вместо СКК-1 для ФКВДио позволяет увеличить ν_2 до 82,9% при $p_0 = 0,1$, при этом P_{ud} уменьшается в 1,23 раза.

Приведенный анализ указывает, что для рассмотренных кодов ФКВД(ио) с СКК-1 и СКК-2 большей эффективностью обладают ФКВД(ио) с СКК-2. Вместе с тем из этого пока не следует вывод в общем случае о меньшей эффективности СКК-1 по сравнению с СКК-2.

ВЫВОДЫ

В процессе проведенного исследования показана возможность факториального кодирования информации, совмещающего функции исправления и обнаружения ошибок, возникающих в канале связи при передаче сообщения. Такое совмещение позволяет повысить динамическую составляющую потери скорости и, как следствие, относительную скорость передачи, по сравнению с обнаруживающим ошибки факториальным кодированием за счет снижения помехоустойчивости кода.

Установлено также, что показатели помехоустойчивости факториального кодирования с восстановлением данных, а также с восстановлением данных и исправлением ошибок не являются инвариантными по отношению к выбору сигнально-кодовой конструкции, если в качестве сигнальных векторов используется некоторое собственное подмножество множества векторов всех возможных перестановок порядка M.

СПИСОК ЛИТЕРАТУРЫ

1. Фауре Э. В. Контроль целостности информации на основе факториальной системы счисления / Э. В. Фауре, В. В. Швыдкий, А. И. Щерба // Journal of Qafqaz University. Mathematics and computer science. – 2016. – № 2. Т. 4. – (В печати).
2. Фауре Э.В. Метод формирования имитовставки на основе перестановок / Э. В. Фауре, В. В. Швыдкий, В. А. Щерба // Захист інформації. – 2014. – №4, Т. 16. – С. 334–340. DOI: 10.18372/2410-7840.16.7620.
3. Фауре Э.В. Комбинированное факториальное кодирование и его свойства / Э. В. Фауре, В. В. Швыдкий, В. А. Щерба // Радіоелектроніка, інформатика, управління. – 2016. – №3. – С. 80–86. DOI: 10.15588/1607-3274-2016-3-10.
4. Фауре Э. В. Факториальное кодирование с восстановлением данных / Э. В. Фауре // Вісник Черкаського державного технологічного університету. – 2016. – № 2. – С. 33–39.
5. Фауре Э. В. Метод повышения эффективности факториального кодирования с восстановлением данных / Э. В. Фауре // Вісник Черкаського державного технологічного університету. – 2016. – №4. – (В печати).
6. Фауре Э. В. Факториальное кодирование с несколькими контрольными суммами / Э. В. Фауре // Вісник Житомирського державного технологічного університету. – 2016. – № 3. – С. 104–113.
7. Питерсон У. Коды, исправляющие ошибки / У. Питерсон, Э. Уэлдон ; пер. с англ. под ред. Р. Л. Добрушина, С. И. Самойленко] – М. : Мир, 1976. – 590 с. – (Редакция литературы по новой технике).
8. Финк Л. М. Теория передачи дискретных сообщений / Л. М. Финк. – Изд. 2-е. – М. : Советское радио, 1970. – 728 с.
9. Теплов Н. Л. Помехоустойчивость систем передачи дискретной информации / Н. Л. Теплов. – М. : Связь, 1964. – 360 с.
10. Конвей Дж. Упаковки шаров, решетки и группы : в 2 т. / Дж. Конвей, Н. Слоэн ; при участии Э. Баннаи и др.; [перевод с англ. С. Н. Лицына и др.] – М. : Мир, 1990. – 2 т.

Статья поступила в редакцию 09.02.2017.
После доработки 25.03.2017.

Фауре Е. В.
Канд. техн. наук, доцент, докторант, доцент кафедри інформаційної безпеки та комп'ютерної інженерії Черкаського державного технологічного університету, Черкаси, Україна

ФАКТОРІАЛЬНЕ КОДУВАННЯ З ВИПРАВЛЕННЯМ ПОМИЛОК

Актуальність. Факторіальне кодування даних дозволяє поєднувати операції крипто- й імітозахисту, а також завадостійкого кодування, що призводить до зменшення внесеної передавачем надлишковості, підвищення швидкодії та збільшення ефективної пропускної здатності. Разом з тим описані методи факторіального кодування не дозволяють виправляти помилки, що обмежує область їх використання.

Метою роботи є розробка методу факторіального кодування з відновленням даних за перестановкою, що забезпечує комплексне вирішення задач криптографічного захисту та завадостійкого кодування і дозволяє поєднати функції виправлення та виявлення помилок каналу зв'язку.

Метод. Основна ідея запропонованого методу кодування полягає в збільшенні відстані між дозволеними кодовими словами, які являють собою перестановки, обчислені за всіма інформаційними бітами блоку даних і представлені в двійковому вигляді. Досліджено методи збільшення відстані на основі метрик Евкліда і Хеммінга. Для кожного з цих методів визначено основні властивості факторіального коду з виправленням помилок, у тому числі виконано оцінку достовірності передавання при незалежності і біноміальному розподілі помилок у каналі зв'язку, розроблено структурні схеми приймача. Правила декодування, реалізовані в приймачі, ґрунтуються на критерії максимальної правдоподібності і передбачають як пряме виправлення помилок, так і їх виявлення з наступним виправленням шляхом перезапиту пошкодженого блоку.

Результати. Реалізовано факторіальні коди з виправленням помилок, які використовують метрики Евкліда і Хеммінга. Для цих кодів виконано порівняльний аналіз ймовірності невиявленої помилки, залишкової ймовірності помилкового прийому, енергетичного виграшу та відносної швидкості передавання. Показано, що характеристики коду не є інваріантними щодо множини дозволених кодових слів, а з розглянутих кодів більш ефективними є коди, які використовують метрику Хеммінга.

Висновки. Отримав подальший розвиток метод факторіального кодування з відновленням даних за перестановкою, який за рахунок поєднання функцій виправлення та виявлення помилок дозволяє підвищити динамічну складову втрати швидкості і, як наслідок, відносну швидкість передавання, в порівнянні з виявляючим помилки факторіальним кодуванням за рахунок зниження його завадостійкості. Проведені експерименти підтвердили ефективність факторіальних кодів з виправленням помилок.

Ключові слова: надлишковість, факторіальний код, перестановка, завадостійке кодування, виправлення помилок, виявлення помилок, достовірність передавання, відносна швидкість передавання.

Faure E. V.

PhD, Associate Professor, Post-Doctoral Associate, Associate Professor of Department of Information Security and Computer Engineering, Cherkasy State Technological University, Cherkasy, Ukraine

FACTORIAL CODING WITH ERROR CORRECTION

Context. Factorial data coding allows combining operations of cryptographic protection, intentional alteration of data, and error-correcting coding which leads to the decrease of redundancy introduced by transmitter and to the increase of data rate and effective throughput. At the same time, the described methods of factorial coding do not correct errors, which limits their use.

Objective of this work is to develop a method of factorial coding with data recovery that provides a comprehensive solution of cryptographic protection and error control coding and allows combining the functions of communication channel errors detecting and correcting.

Method. The basic idea of the proposed coding method is to increase the distance between the allowed code words that represent permutations calculated for all information bits of a data block and represented in a binary form. The methods of distance increasing based on Euclidean and Hamming metrics are investigated. The basic properties of factorial code with error correction are defined for each of these methods. The estimate of probability characteristics is done on the condition of independence of communication channel errors and their binomial distribution. The receiver structures are developed. Decoding rules implemented in receiver are based on the maximum likelihood criteria and provide both forward error correction and error detection with further correction by retransmission of damaged data block.

Results. The factorial error-correcting codes using Euclidean and Hamming metrics are implemented. The comparative analysis of the probability of an undetected error, the residual probability of erroneous reception, energy gain, and the relative transmission rate is done for these codes. It is shown that code characteristics are not invariant to the set of allowed code words, and the codes that use Hamming metric are the most efficient codes between the presented codes.

Conclusions. The method of factorial coding data recovery by permutation has been further developed. Due to the combination of error correction and detection functions, it can increase the rate loss dynamic component and, consequently, the relative transmission rate, compared to error-detecting factorial coding by reducing its noise immunity. The experiments confirmed the effectiveness of the factorial error-correcting codes.

Keywords: redundancy, factorial code, permutation, error control coding, error correction, error detection, reliability of data transmission, relative transmission rate.

REFERENCES

1. Faure E'. V., Shvydkij V. V., Shherba A. I. Kontrol' celostnosti informacii na osnove faktorial'noj sistemy schisleniya, *Journal of Qafqaz University. Mathematics and computer science*, 2016, No. 2, Vol. 4. (V pechati).
2. Faure E'. V., Shvydkij V. V., Shherba V. A. Metod formirovaniya imitovstavki na osnove perestanovok, *Zaxist informacii*, 2014, No. 4, Vol. 16, pp. 334–340. DOI: 10.18372/2410-7840.16.7620.
3. Faure E'.V., Shvydkij V. V., Shherba V. A. Kombinirovannoe faktorial'noe kodirovanie i ego svojstva, *Radio Electronics, Computer Science, Control*, 2016, No. 3, pp. 80–86. DOI: 10.15588/1607-3274-2016-3-10.
4. Faure E'.V. Faktorial'noe kodirovanie s vosstanovleniem dannyx, *Visnyk Cherkas'kogo derzhavnogo tehnologichnogo universytetu*, 2016, No. 2, pp. 33–39.
5. Faure E'.V. Metod povysheniya e'ffektivnosti faktorial'nogo kodirovaniya s vosstanovleniem dannyx, *Visnyk Cherkas'kogo derzhavnogo tehnologichnogo universytetu*, 2016, No. 4. (V pechati).
6. Faure E'.V. Faktorial'noe kodirovanie s neskol'kimi kontrol'nymi summami, *Visnyk Zhytomyrs'kogo derzhavnogo tehnologichnogo universytetu*, 2016, No. 3, pp. 104–113.
7. Piterson U., Ue'ldon E'.; [per. s angl. pod red. R. L. Dobrushina, S. I. Samojlenko] Kody, ispravlyayushhie oshibki. Moscow, Mir, 1976, 590 p. (Redakciya literatury po novoj texnike).
8. Fink L. M. Teoriya peredachi diskretnyx soobshhenij. [Izd. 2-e, pererab. i dopoln.]. Moscow, Sovetskoe radio, 1970, 728 p.
9. Teplov N. L. Pomexoustojchivost' sistem peredachi diskretnoj informacii. Moscow, Svyaz', 1964, 360 p.
10. Konvej Dzh., Sloe'n N. ; pri uchastii E'. Bannai i dr.; [perevod s angl. S. N. Licyna i dr.] Upakovki sharov, reshetki i gruppy : v 2 t. Moscow, Mir, 1990, 2 t.

UDC 004.272.26: 004.93

Oliinyk A.[1], Subbotin S.[2], Skrupsky S.[3], Lovkin V.[4], Zaiko T.[5]

[1]PhD., Associate Professor of Department of Software Tools, Zaporizhzhia National Technical University, Zaporizhzhia, Ukraine
[2]Dr.Sc, Head of Department of Software Tools, Zaporizhzhia National Technical University, Zaporizhzhia, Ukraine
[3]PhD, Associate Professor of Computer Systems and Networks Department, Zaporizhzhia National Technical University, Zaporizhzhia, Ukraine
[4]PhD, Associate Professor of Department of Software Tools, Zaporizhzhia National Technical University, Zaporizhzhia, Ukraine
[5]PhD., Senior Lecture of Department of Software Tools, Zaporizhzhia National Technical University, Zaporizhzhia, Ukraine

INFORMATION TECHNOLOGY OF DIAGNOSIS MODELS SYNTHESIS BASED ON PARALLEL COMPUTING

Context. The problem of diagnosis models synthesis in the big data processing based on parallel computing is solved. The object of the research is the process of diagnosis models synthesis. The subject of the research are the methods and information technologies for diagnosis models synthesis.

Objective. The research objective is to develop diagnosis models synthesis information technology.

Method. The paper deals with information technology of diagnosis models synthesis which is a set of diagrams graphically describing structural elements of the system as well as the behavioral aspects of their interaction at various stages of diagnostics objects models construction. The developed information technology enables to perform the construction of distributed diagnostics systems where computationally complex stages of diagnosis models synthesis are performed on high-performance server equipment, which makes it possible to significantly increase the practical threshold for using diagnostics systems in the processing of big data sets for solving of the tasks of training sample data reduction, rules extraction, diagnosis models construction and retraining.

Results. The software which implements the proposed information technology and allows to synthesize diagnosis models based on the given data samples has been developed.

Conclusions. The conducted experiments have confirmed the proposed information technology operability and allow to recommend it for solving the problems of big data processing for technical and biomedical diagnostics in practice. The prospects for further researches may include the modification of the developed information technology by introducing of other methods of diagnosis models synthesis.

Keywords: data sample, diagnosis, rule extraction, feature selection, parallel computing, model synthesis.

NOMENCLATURE

CPU – Central Processing Unit;
GPU – Graphical Processing Unit;
IGA – Island Genetic Algorithm;
UML – Unified Modeling Language;
E_{IGA} – efficiency of the computer system in the implementation of the IGA method;
E_p – efficiency of the computer system in the implementation of the proposed method;
M – number of features in the sample of observations S;
$Overhead_{IGA}$ – communication and synchronization costs of the IGA method;
$Overhead_p$ – communication and synchronization costs of the proposed method;
P – set of features (attributes) of observations in the given sample;
Q – number of observations in the given sample of observations S;
S – sample of observations (training sample);
S_{IGA} – speedup of the IGA method;
S_p – speedup of the proposed method;
T – set of output parameter values;
T_{IGA} – execution time of the IGA method;
T_p – execution time of the proposed method.

INTRODUCTION

Solving of technical and medical diagnostics tasks, as well as the task of nondestructive product quality control, is connected with the task of diagnosis models construction [1–6]. Such models enable to classify samples which are diagnosed with high accuracy, and at that the target classes are sufficiently convenient for perception and analysis by experts in the application areas.

Methods which enable to solve training sample data reduction, rules extraction, diagnosis models detection and retraining problems are presented in the papers [7–16]. However such methods based on sequential computing require significant time costs, which makes it difficult to use such methods for diagnostic decision making process automation in practice. So it is necessary to develop information technology which is capable to accelerate diagnosis models synthesis process based on parallel computing [17–20].

The research objective is to develop diagnosis models synthesis information technology which enables to build distributed diagnostics systems where computationally complex stages of model synthesis are performed on high-performance server equipment, which makes it possible to significantly increase the practical threshold for using diagnostic systems in big data sets processing.

1 PROBLEM STATEMENT

Suppose we have data sample $S = < P, T >$, which consists of Q samples. Every sample is characterized by values of attributes p_{q1}, p_{q2}, …, p_{qM} and output parameter t_q, where p_{qm} is a value of the m-th attribute of the q-th sample, $q = 1, 2, ..., Q$, $m = 1, 2, ..., M$; M – overall number of attributes in the set S. Then the problem of diagnosis models synthesis is to define model DM for

DOI 10.15588/1607-3274-2017-3-16

the sample $S =< P,T >$, which allows to approximate the set S with acceptable error level E. Error E [2, 21–23] of recognition using synthesized model DM can be evaluated as ratio of incorrectly recognized samples number N_{er} to overall samples number Q in the sample $S =< P,T >$: $E = \frac{N_{er}}{Q}$.

2 REVIEW OF THE LITERATURE

As stated above, in the papers [7–16] methods, which enable to solve problems concerning diagnosis models synthesis, are proposed. However such methods based on sequential computing require significant time costs which makes it difficult to use such methods for diagnostic decision making process automation in practice.

Feature selection methods, which were proposed in the papers [7–9], enable to select informative feature combinations based on evolutionary [24–26] and multi-agent technologies [27] of computational intelligence. The proposed methods use aprioristic information about individual self-descriptiveness, reducing search space and decreasing time of informative feature combination selection. Nevertheless such methods require significant time costs for implementation during processing of high dimensionality data.

Productional rules extraction methods [10–12] enable to find the most significant replications $X \to Y$ from the given data samples $S =< P,T >$. It provides improvement of cumulative properties of diagnosis models which are synthesized, as well as increases interoperability of models, decreases its dimensionality (structural and parametrical complexity), utilized storage capacity and response speed.

The method of parametrical identification of neuro-fuzzy networks based on parallel random search, which was proposed in [13–16], uses stochastic optimization for synthesized models parameters setting (parameters of membership functions and weight coefficients of neuroelements), forms initial solution set subject to training sample information (significance of feature terms according to the compactness of training set samples arrangement in the corresponding term and the degree of its effect on output parameter value). It makes it possible to move initial search points near optimal values and to accelerate optimization process.

It is necessary to develop information technology which enables to use the methods proposed in the papers [7–16] in practice for construction of distributed diagnostics systems where computationally complex stages of diagnosis models synthesis are performed on high-performance server equipment, significantly increasing the practical threshold for using diagnostic systems.

3 MATERIALS AND METHODS

As it was mentioned above, diagnosis models construction based on available data sets generally requires considerable computational resources. Therefore the developed information technology is implemented based on «client-server» architecture [28–30]. At that it is proposed to implement complex computational processes concerning training sample processing, big data storing, model synthesis etc., on server side, organizing client access according to their access permissions. Clients are understood as people and computer systems which solve practical tasks concerning construction and application of diagnosis models.

The diagram describing main system functions of information technology front-end and back-end is presented in Fig. 1.

As it is described in Fig. 1, main functions of information system are divided between server and client sides. Client side of information system provides implementation of user interface, processing of data, entered by user (verification of data format correctness), and calculation of diagnostics object output parameter value based on the synthesized model. Server side of the system should use high-performance equipment for solving of the tasks concerning diagnosis models construction. Computationally complex processes concerning reduction of training sample dimensionality, rules extraction, synthesis and retraining of diagnosis models are processed on the server side. Besides it is proposed to use database for storage and processing of training samples from different users and also of synthesized models, extracted rules, reduced samples and other results of system operation.

For the design and development possibility of diagnosis models synthesis software system, corresponding information technology presented as the set of UML-models [31–33] is proposed. UML-models are represented as diagrams (Fig. 2–7), graphically describing structural and behavioral aspects of construction of distributed diagnostics systems, which enable to solve the tasks of training sample data reduction, rules extraction, diagnosis models construction and retraining.

Based on the main functions of the system (Fig. 1) and also on the chosen architecture of diagnosis models synthesis information technology, it is possible to define general configuration and topology of the system as the model presented in deployment diagram (Fig. 2).

System consists of three nodes: server, client computer and database server. Software implementation of intellectual methods which are used at various stages of diagnosis models synthesis (data reduction, rules extraction, model construction, model retraining) is contained on the server. Set of client computers with installed software can interact with the server. Work of database server is organized through interaction between database management system and database. Interaction of user (client computers) with database is performed indirectly through the server, where user access permissions for corresponding data sets are verified.

For representation of different user interactions with the system the model was figured as use case diagram (Fig. 3).

As can be seen from the Fig. 3, there are two user kinds which interact with the system: user and administrator. Interaction between them is realized in the use case “Registration”, when administrator approves user permissions for access to the other system functions.

System user is a person which uses information technology for solving practical diagnostics tasks. User should be registered in the system to provide access of

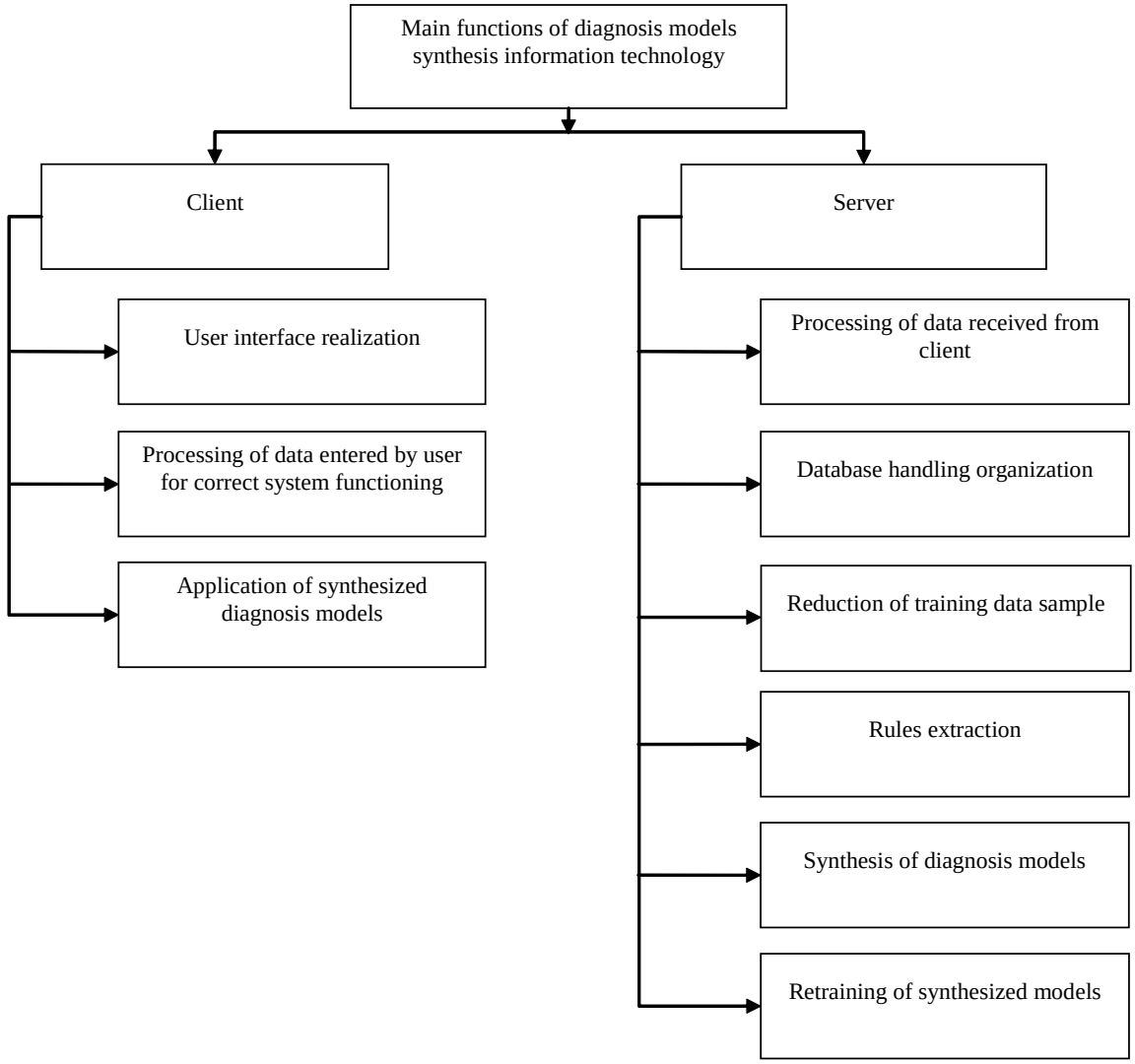


Figure 1 – Main functions of diagnosis models synthesis information technology

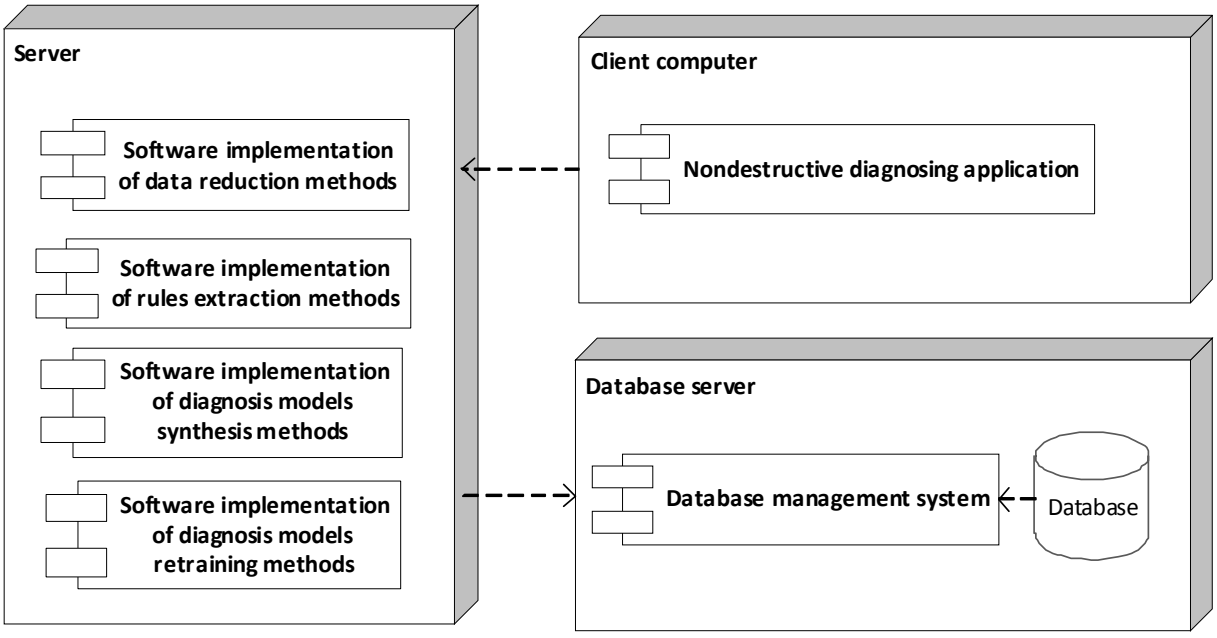


Figure 2 – Deployment diagram of diagnosis models synthesis information technology (UML 2.5 notation)

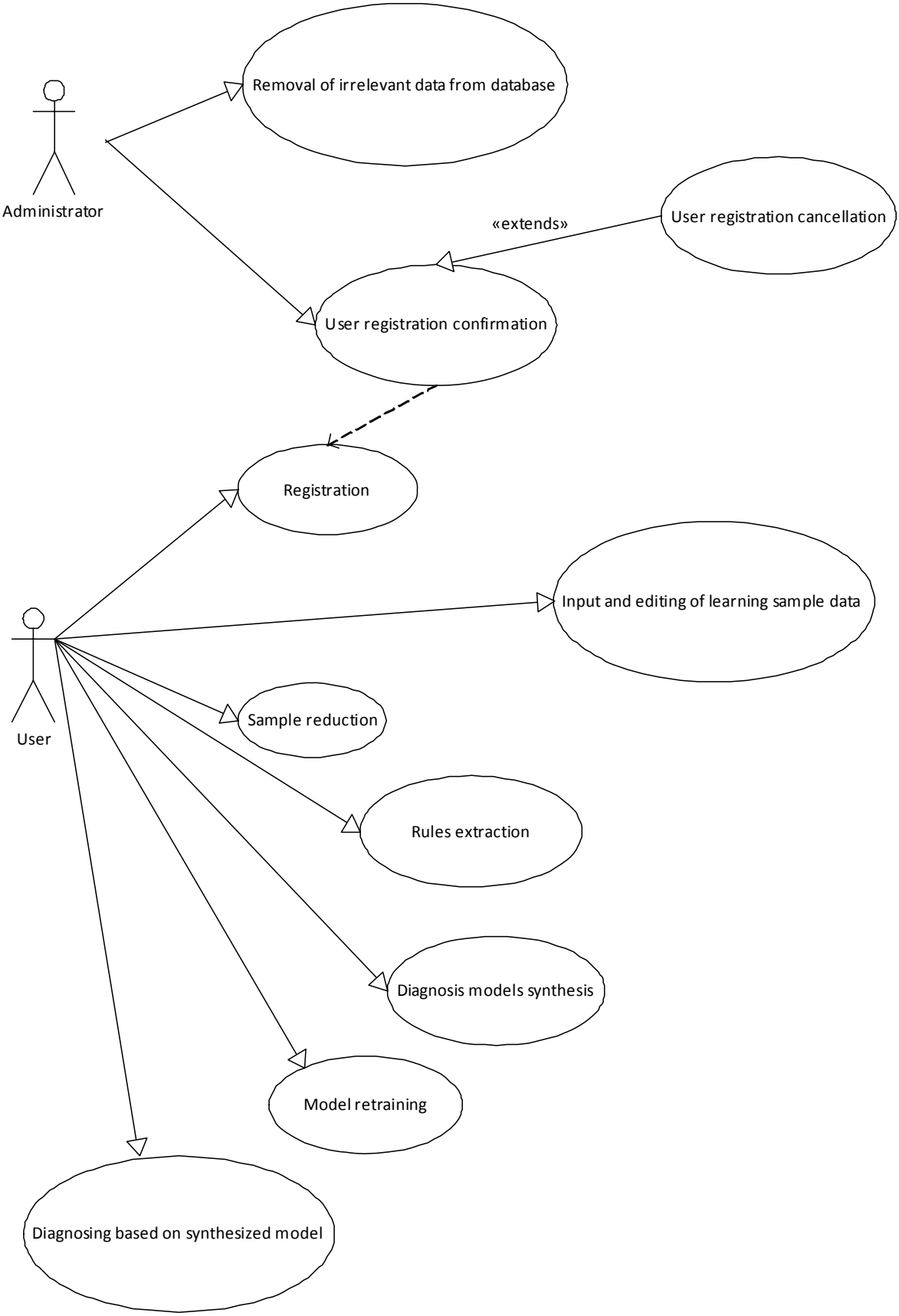


Figure 3 – Use case diagram of diagnosis models synthesis information technology

different users to the corresponding data samples. User can interact with the system during registration, input and editing of training sample, solving of diagnostics tasks (training sample dimensionality reduction, extraction of new knowledge about the dependencies which are researched, synthesis or diagnostics model retraining, diagnostics based on the synthesized model).

System administrator provides user registration through input and editing of information about users and user registration confirmation, and also provides access to the database for removal or archiving of information which is not used during long period of time and reduction in such a way of physical volume of the disk space which is in use.

With the object of modeling of logical operation and actions performed in the diagnosis models synthesis information system, the corresponding model was created and is presented in the Fig. 4 as activity diagram.

As it is presented in the Fig. 4, user enters account data at the beginning of system usage. After that system proposes to choose one of the following operating modes:

– diagnosis model synthesis – computationally complex process, which is realized on server, performing stages of training sample data reduction, rules extraction, diagnosis model construction;

– retraining of the synthesized model;

– diagnosing using the synthesized model – calculation of diagnostics object output parameter value using the model based on the measured input parameters.

Then user can continue to use the system or can quite.

The model of distribution of interaction between information system objects and users in time is represented as sequence diagram (Fig. 5–7). In the diagram the process of interaction between system components through calls of procedures, which realize corresponding use cases, is presented.

In the sequence diagram, presented in the Fig. 5, system use cases concerning user registration, input, editing of training sample and removing of irrelevant data (Fig. 3) are represented. As can be seen, not only users but also system administrator takes part in these processes.

Processes connected with preliminary processing of training sample for diagnosis model synthesis (reduction of data sample and rules extraction) are performed on user request (through client computer) on the server, and the results are saved in the database (Fig. 6). Processes, presented in the sequence diagram (Fig. 6), correspond to the system use cases “Sample reduction” and “Rules extraction” (Fig. 3), which are performed by user without administrator participation.

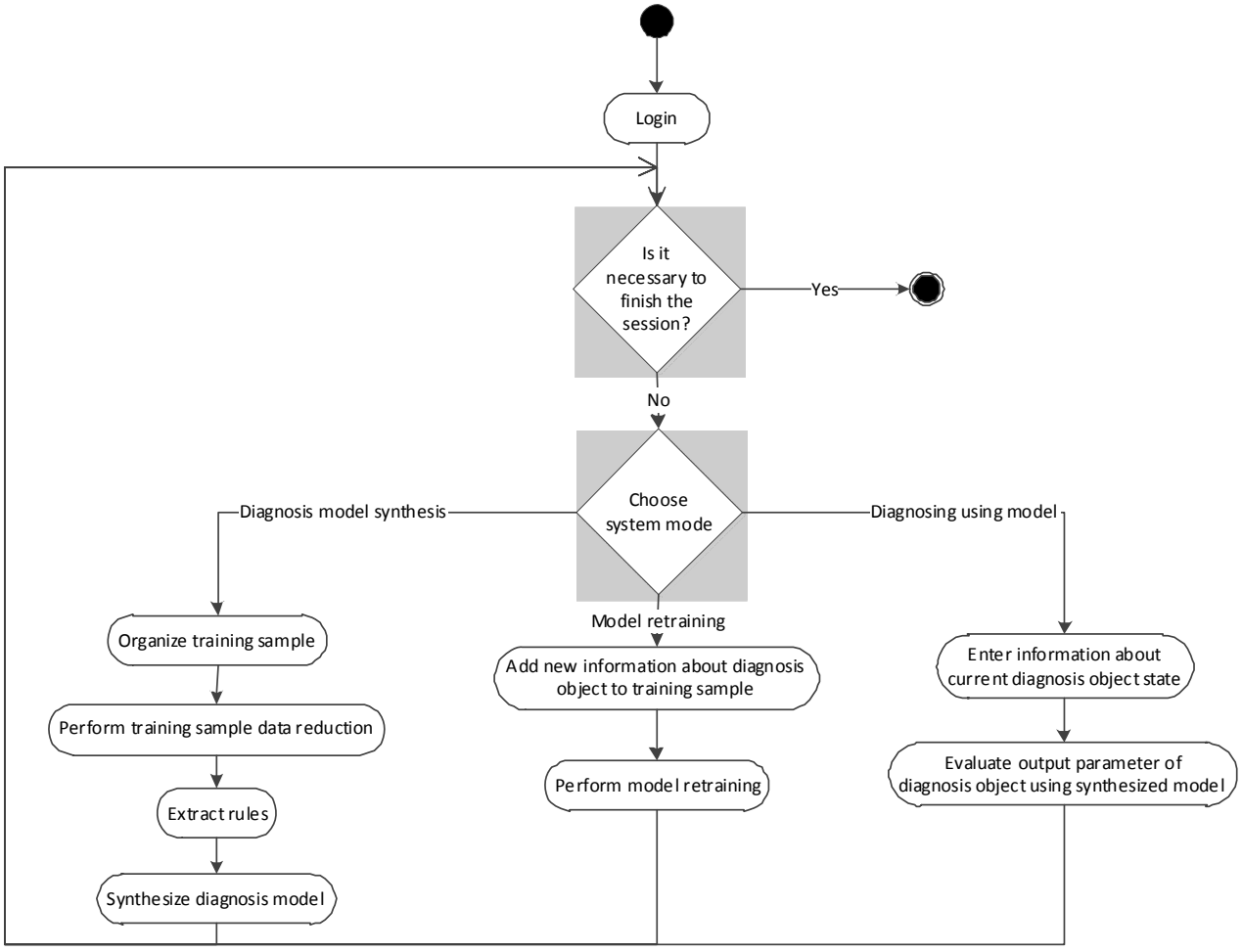


Figure 4 – Activity diagram of diagnosis models synthesis information technology

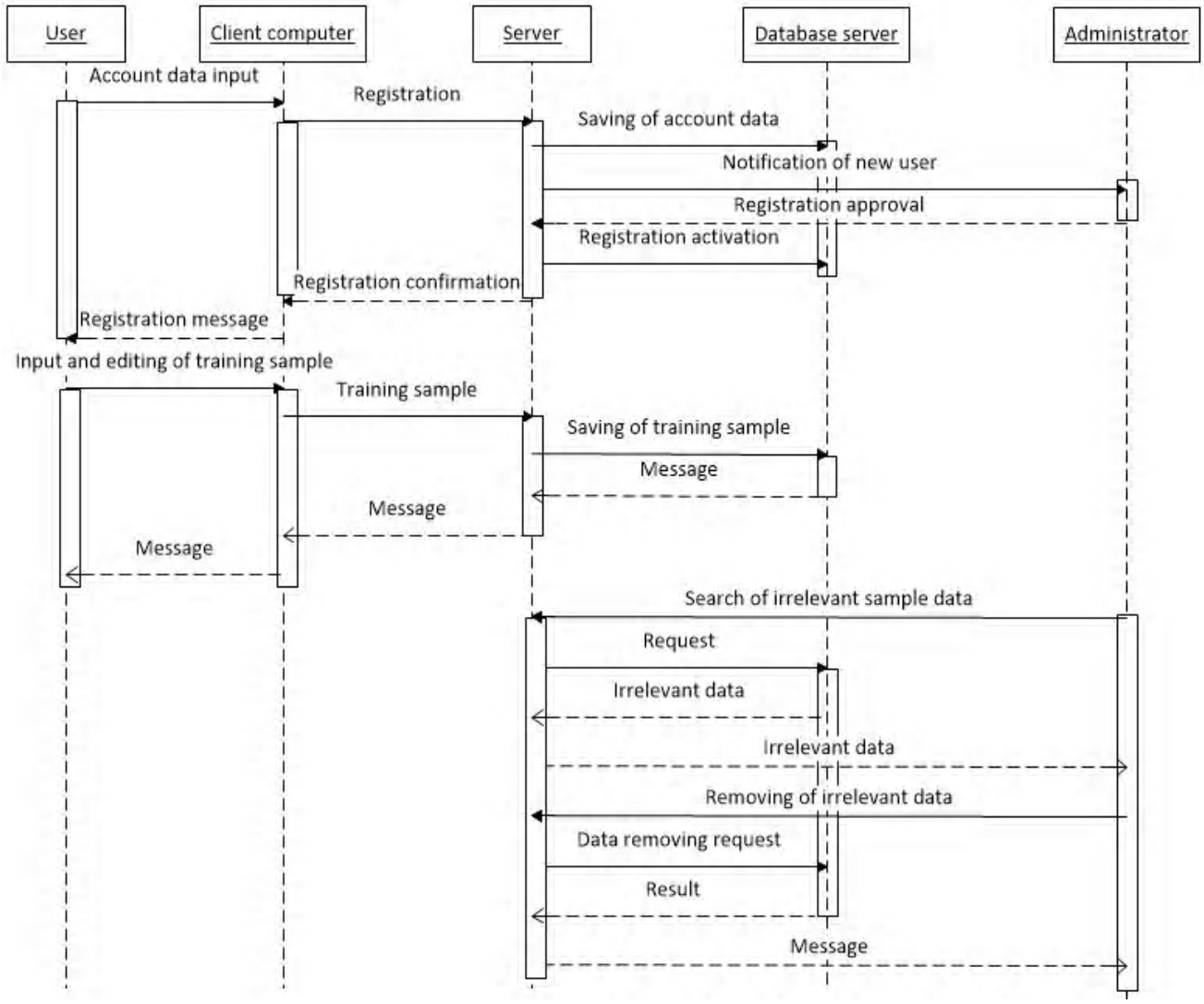


Figure 5 – Sequence diagram of user registration and training sample processing

Computationally complex processes of model synthesis and retraining are also performed on the server on user request (Fig. 7) with saving of results to the database. The result of these processes (synthesized diagnosis model) is transmitted to client computer. It enables further usage of the synthesized model for diagnosis object or process output parameter value calculation. Such approach provides diagnostics process execution on client computer without server access.

Thus the developed information technology of diagnosis models synthesis is represented by the set of diagrams (Fig. 3–8), which graphically describe structural elements of the system, and also behavioral aspects of its interaction at various stages of diagnostics objects models construction. The proposed information technology allows to construct distributed diagnostics systems, where computationally complex stages of diagnosis models synthesis are performed on the high-performance server equipment, which enables to significantly increase the practical threshold for using diagnostic systems which are capable of solving the tasks of training sample data reduction, rules extraction, diagnosis models construction and retraining.

As it was mentioned above, it is supposed to use database for storage and processing of information about objects or processes which are researched (training samples) and also for system operation results (synthesized models, extracted rules, reduced samples etc.) in the developed diagnosis models synthesis information technology. Database contains set of tables, which are connected in some way and contain information about users and samples of data representing objects and processes which are researched [34–36]. ER-model of the developed diagnosis models synthesis information technology database supposes availability of the following entities:

– Users – contains information about system users. Entity fields: id – user number (unique identifier); userLogin – login of user; userPass – password of user; info – information about user;

– Groups – reflects user groups and group data. Fields: id – primary key, name – group name, info – group description;

– UserGroups – describes correspondence between users and their groups. At that idGroup, idUser are foreign keys pointing at the corresponding entities;

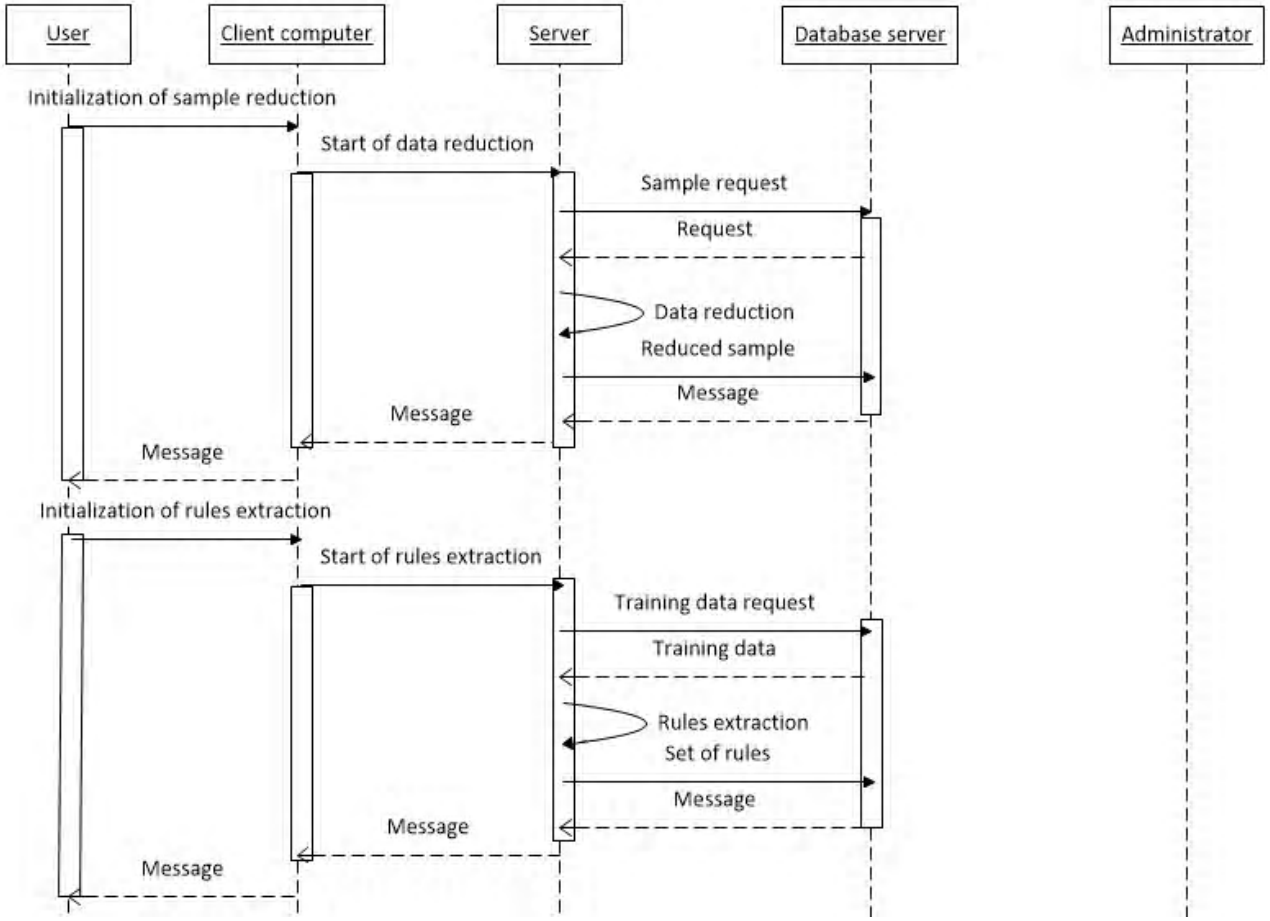


Figure 6 – Sequence diagram of training sample reduction and rules extraction

– Samples – contains information about training samples. Entity fields: id – unique identifier of the corresponding data sample; ref – reference to the file with the table which contains training sample $S = < P,T >$ with the identifier id; info – description of training sample; idGroups – reference to the user group which has permissions for access to the sample id;

– SamplesRed – contains information about the samples which were reduced using the methods implemented in the system through the reduction of training samples from the entity Samples. Fields: id – unique identifier of reduced data sample; ref – reference to the file with the table which contains reduced sample, corresponding to the identifier id; info – description of reduced sample (particularly reduction method); idSample – identifier (foreign key) of the training sample which was the base for getting of the reduced sample;

– Rules – contains information about productional rules, which were extracted using methods implemented in the system through the processing of training samples from the entity Samples. Entity fields: id – unique identifier of rules set; ref – reference to the file which contains set of rules $P \to T$, corresponding to the identifier id; info – description of rules set (particularly rules extraction method); idSample – identifier (foreign key) of the training sample $S = < P,T >$, which was the base for getting of the set of rules $P \to T$;

– Models – contains information about diagnosis models, which were synthesized based on the given (idSample) or reduced (idSampleRed) samples, and also on the extracted rules (idRules) using the methods which were implemented in the system. Fields: id – unique identifier of diagnosis model; ref – reference to the file which contains structure and parameters of the model with identifier id; info – description of the synthesized model (for example, synthesis method); idSample, idSampleRed, idRules – identifiers (foreign keys) of the training sample, reduced sample and rules set correspondingly, which were the base for the synthesis of the model id. At the same time if only one data set is used in the synthesis of the model id, it is necessary to use reference to the dummy record in the other entities Samples, SamplesRed, Rules as values of the other foreign keys; idParentModel – model which was used as the base for synthesis of the model id (this parameter is not required, because it is necessary only if synthesized model is retrained); idGroups – reference to the group of users having access to the implementation of the synthesized model.

Database scheme is presented in the Fig. 8.

As it is presented in the Fig. 8, connections between entities are created from foreign to primary key, and integrity control is provided by database management system tools.

The developed database supports storage and processing of information about objects and processes which are researched (training samples) and also system

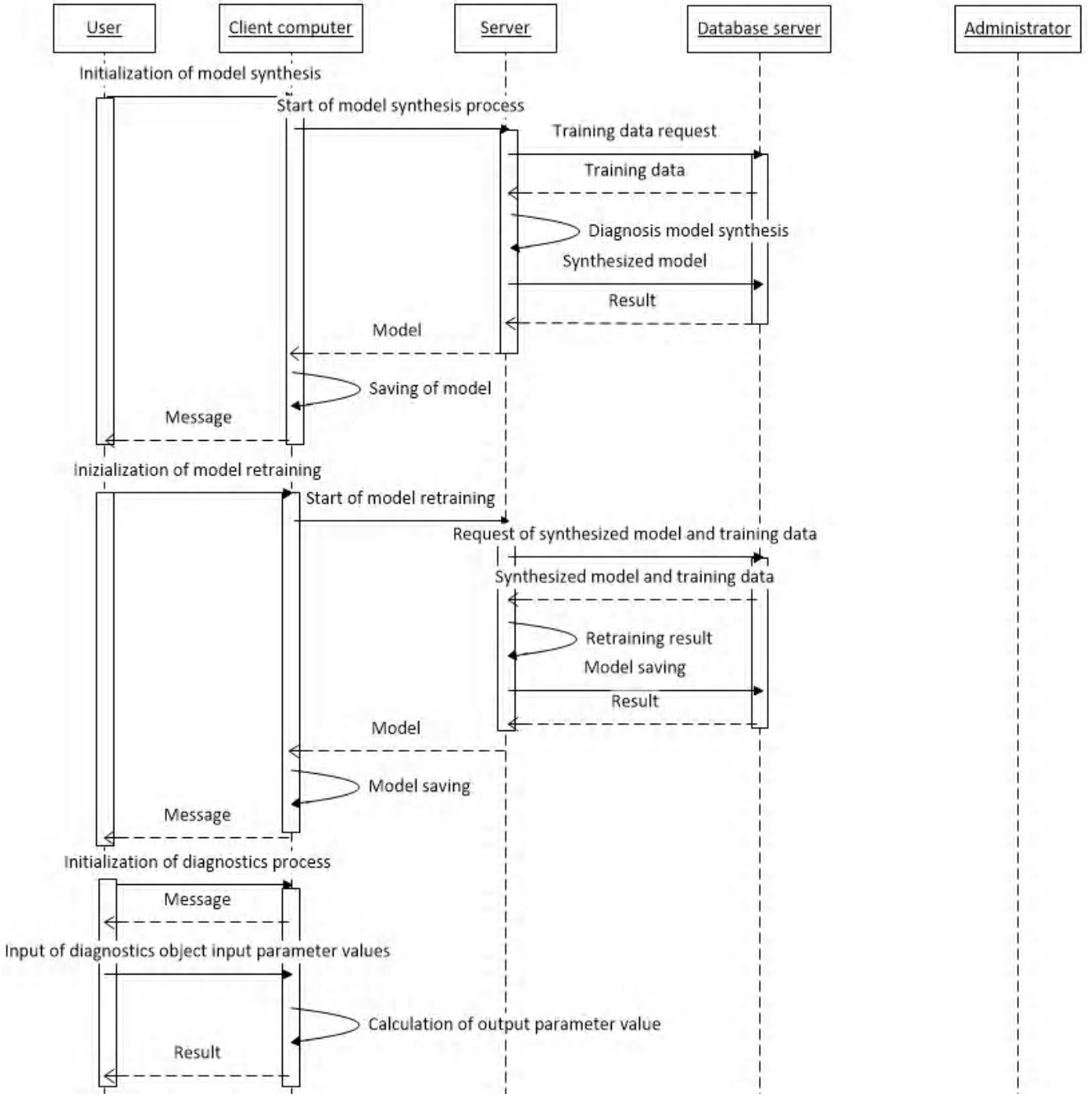


Figure 7 – Sequence diagram of synthesis and usage of diagnosis models

operation results (structure and parameters of synthesized models, extracted rules, reduced samples etc.) in the developed diagnosis models synthesis information technology.

4 EXPERIMENTS

For efficiency examination of the proposed information technology corresponding software system for diagnosis models synthesis was developed.

The software was developed based on Java programming language and architecture pattern MVC. Graphic part (View) was implemented using SWING package. The application has the following class structure. Class Model represents classes which describe entities of database ER-model (Fig. 8). Fields of these classes are identical with fields of database tables. It enables to apply modern programming frameworks, for example, Hibernate. Every row of database table can be got as one class instance in the application. At that several table rows form collection. Classes View are frames (forms, dialog boxes), which contain graphical user interface and allow user to interact with the application. Classes Controller connect to the server, realize event service and handling of user actions. Class LoginController enables to connect to the server using login and password or to register (to create user account which is inserted to the database). MainController enables to handle user instructions of calling appropriate forms. It sends instruction to the server and gets permission (according to

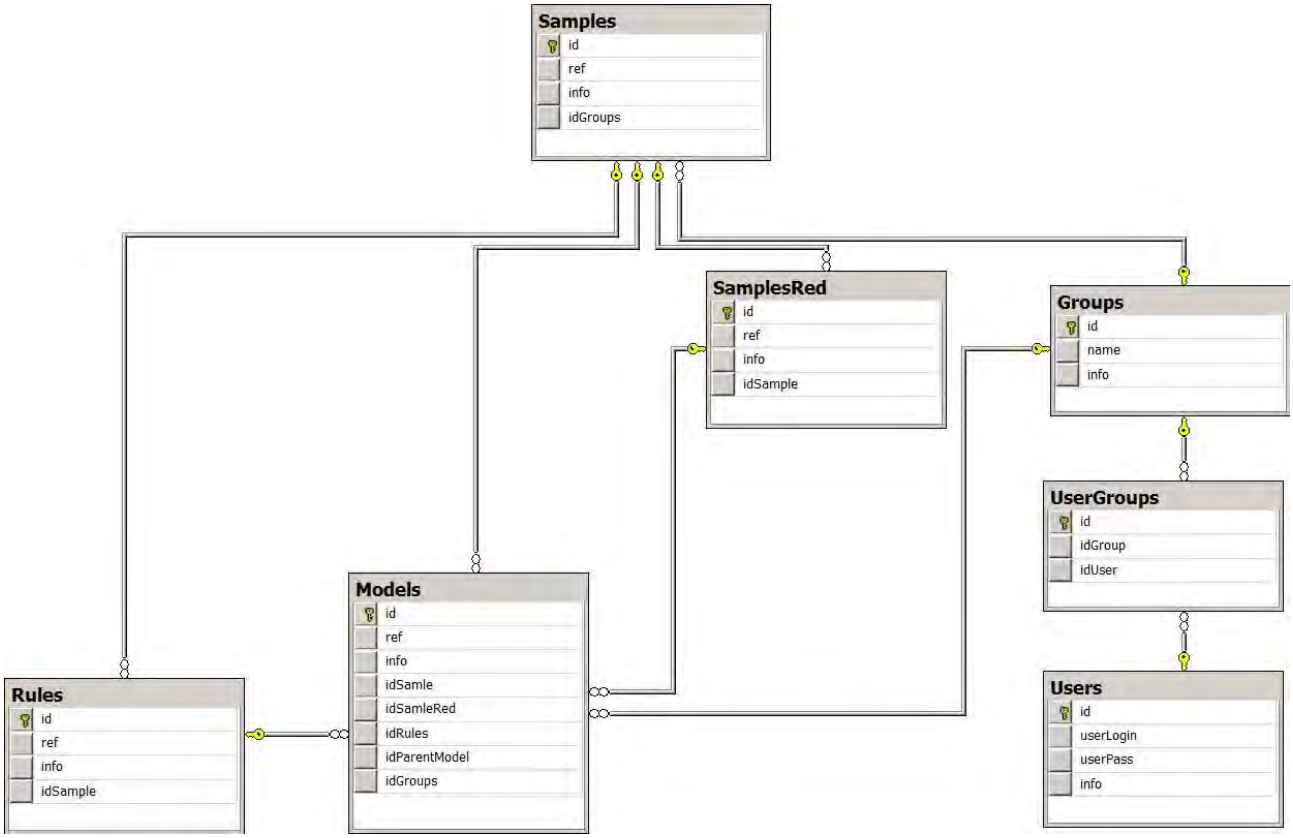


Figure 8 – Diagnosis models synthesis information technology database scheme

user access permissions) to call appropriate function, for example, diagnostics or synthesis. Class WorkWithSampleController gives methods, which enable to load sample from the file with the following sending of it to the server through ConnectionController. Besides it enables to edit sample and to reduce it. Class SynthesisController downloads samples and synthesis methods which are available for user from database and also performs models synthesis. Class AdditionalTrainingController gives methods, which get models and retraining methods available for user from database, besides it enables to extend sample and to retrain model. DiagnosisController enables to access diagnosis models, to input model parameters and to diagnose with the following saving of the result to the file.

Numerical experiments on the efficiency of the proposed information technology and the corresponding software system were performed by solving of different practical diagnostics problems [37–39], particularly the task of hypertensive patient health status prediction [37]. Applying the developed mathematical (methods [7–16] and the proposed information technology) and software support the tasks of learning sample reduction (informative attributes identification and production rules extraction) and diagnosis models synthesis were solved sequentially. Numerical results of the developed information technology application for neuro-fuzzy diagnosis models definition using the proposed parallel method based on stochastic computation [24–27, 40] and island method of evolutional search (Island Genetic Algorithm, IGA) [24–26] are presented. Experiments were executed on 1, 2, 4, 8 and 16 cores of CPU cluster [41] as well as on graphic engine GPU [42].

5 RESULTS

Experimental results using CPU cluster are presented in the table 1.

The results of experiments using graphic engine GPU NVIDIA GTX 285+, which was programmed based on CUDA technology [43], are presented in the table 2. Speedup of computational process was measured regarding one CPU.

Table 1 – The results of experiments using CPU cluster

CPU number	T_p, s	T_{IGA}, s	S_p	S_{IGA}	$Overhead_p$, s	$Overhead_{IGA}$, s	E_p	E_{IGA}
1	101.18	128.5	1	1	0	0	1	1
2	57.67	74.53	1.75	1.72	8.07	11.92	0.88	0.86
4	30.61	39.19	3.31	3.28	6.43	8.62	0.83	0.82
8	18.09	23.61	5.59	5.44	7.78	11.1	0.7	0.68
16	11.45	15.34	8.84	8.38	9.27	13.96	0.55	0.52

Table 2 – The results of experiments realized on GPU

Number of GPU threads	T_p, s	T_{IGA}, s	S_p	S_{IGA}	$Overhead_p$, s	$Overhead_{IGA}$, s
60	39.8	60.22	2.54	2.13	4.46	7.05
80	33.49	50.86	3.02	2.53	4.09	6.41
100	30.33	46.18	3.34	2.78	4.0	6.23
120	28.89	44.25	3.5	2.9	4.13	6.64
140	28.55	44.7	3.54	2.87	4.37	6.93
160	29.05	45.72	3.48	2.81	4.74	7.73
180	30.29	47.96	3.34	2.68	5.24	8.63
200	31.0	48.92	3.26	2.63	5.7	9.15
240	31.26	51.37	3.24	2.5	6.56	13.36

6 DISCUSSION

As it is presented in the tables 1 and 2, the proposed technology of diagnosis models synthesis allows to synthesize models with productivity similar to the models described in [3, 10, 13]. Thus, the method proposed in [13] due to application of new solution search operators modifications decreases number of processor operations, including communicatory costs, and so random search is realized quicker than in the IGA method [24–26] (for example, time of model synthesis for 16 cores of CPU equaled 11.45 s for the proposed method and 15.34 s for IGA method). At that the proposed diagnosis models synthesis technology provides construction of neuro-fuzzy models with acceptable accuracy. That is productivity rise is not provided due to decrease of diagnosis models approximating and resumptive properties level.

The efficiency of CPU cluster, which was used by the method proposed in [13] and the IGA method, is acceptable (particularly parallel system efficiency reaches 0.7 for the proposed method and 0.68 for the IGA method using 8 cores of CPU). Application of more than 8 cores of CPU isn't justified, because it greatly decreases system efficiency due to transmission and synchronization. If number of GPU threads rises above 140, speedup of computing process will decrease, because overheads considerably rise and at the same time threads begin to stand.

Thus diagnosis models synthesis technology which was proposed in the paper allows to efficiently apply modern parallel computing architectures for getting the result with appropriate accuracy in acceptable time. Usage of cross-platform language considerably extends scope of the proposed technology.

CONCLUSIONS

In this paper actual problem of diagnosis models synthesis process automation was solved.

Scientific novelty of the paper is in the proposed information technology of diagnosis models synthesis, which consists of the set of methods and diagrams which connect methods with each other, graphically describe structural elements of diagnostics systems and also behavioral aspects of its communication at various stages of diagnostics objects models construction. The proposed information technology enables to construct distributed diagnostics systems where computationally complex stages of diagnosis model synthesis are performed on high-performance server equipment, which makes it possible to significantly increase the practical threshold for using diagnostic systems in big data sets processing, solving the tasks of training sample data reduction, rules extraction, building and retraining of diagnosis models.

Practical significance of the paper consists in the solution of practical problems. Experimental results showed that the proposed information technology allowed to significantly increase the speed of diagnosis models synthesis process and it could be used in practice for solving of practical tasks concerning diagnostics and nondestructive product quality control.

ACKNOWLEDGMENTS

The work was performed as part of research work "Methods and means of computational intelligence and parallel computing for processing large amounts of data in diagnostic systems" (number of state registration 0116U007419) of software tools department of Zaporizhzhia National Technical University with partial support of international project "Internet of Things: Emerging Curriculum for Industry and Human Applications" (ALIOT, Ref. No. 573818-EPP-1-2016-1-UK-EPPKA2-CBHE-JP) co-funded by the Erasmus+ programme of the European Union.

REFERENCES

1. Price C. Computer based diagnostic systems / C. Price. – London : Springer, 1999. – 136 p. DOI: 10.1007/978-1-4471-0535-0.
2. Bow S. Pattern recognition and image preprocessing / S. Bow. – New York : Marcel Dekker Inc., 2002. – 698 p. DOI: 10.1201/9780203903896.
3. Encyclopedia of machine learning / [eds. C. Sammut, G. I. Webb]. – New York: Springer, 2011. – 1031 p. DOI: 10.1007/978-0-387-30164-8.
4. Bezdek J. C. Pattern Recognition with Fuzzy Objective Function Algorithms / J. C. Bezdek. – N.Y. : Plenum Press, 1981. – 272 p. DOI: 10.1007/978-1-4757-0450-1.
5. Sobhani-Tehrani E. Fault diagnosis of nonlinear systems using a hybrid approach / E. Sobhani-Tehrani, K. Khorasani. – New York: Springer, 2009. – 265 p. – (Lecture notes in control and information sciences ; № 383). DOI: 10.1007/978-0-387-92907-1.
6. Bodyanskiy Ye. Hybrid adaptive wavelet-neuro-fuzzy system for chaotic time series identification / Ye. Bodyanskiy, O. Vynokurova // Information Sciences. – 2013. – Vol. 220. – P. 170–179. DOI: 10.1016/j.ins.2012.07.044.
7. Subbotin S. The Dimensionality Reduction Methods Based on Computational Intelligence in Problems of Object Classification and Diagnosis / S. Subbotin, A. Oliinyk // Recent Advances in Systems, Control and Information Technology. Advances in Intelligent Systems and Computing. – 2017. – Vol. 543. – P. 11–19. DOI: 10.1007/978-3-319-48923-0_2.
8. Subbotin S. The Sample and Instance Selection for Data Dimensionality Reduction / S. Subbotin, A. Oliinyk // Recent Advances in Systems, Control and Information Technology. Advances in Intelligent Systems and Computing. – 2017. – Vol. 543. – P. 97–103. DOI: 10.1007/978-3-319-48923-0_13.

9. Oliinyk A. A. Parallel multiagent method of big data reduction for pattern recognition / A. A. Oliinyk, S. Yu. Skrupsky, V. V. Shkarupylo, O. Blagodariov // Radio Electronics, Computer Science, Control. – 2017. – № 2. – С. 82–92.
10. Oliinyk A. Production rules extraction based on negative selection / A. Oliinyk // Radio Electronics, Computer Science, Control. – 2016. – № 1. – P. 40–49. DOI: 10.15588/1607-3274-2016-1-5.
11. Oliinyk A. The decision tree construction based on a stochastic search for the neuro-fuzzy network synthesis / A. Oliinyk, S. A. Subbotin // Optical Memory and Neural Networks (Information Optics). – 2015. – Vol. 24, № 1. – P. 18–27. DOI: 10.3103/S1060992X15010038.
12. Oliinyk A. A stochastic approach for association rule extraction / A. Oliinyk, S. A. Subbotin // Pattern Recognition and Image Analysis. – 2016. – Vol. 26, № 2. – P. 419–426. DOI: 10.1134/S1054661816020139.
13. Oliinyk A. O. Using Parallel Random Search to Train Fuzzy Neural Networks / A. O. Oliinyk, S. Yu. Skrupsky, S. A. Subbotin // Automatic Control and Computer Sciences. – 2014. – Vol. 48, Issue 6. – P. 313–323. DOI: 10.3103/S0146411614060078.
14. Oliinyk A. Parallel computing system resources planning for neuro-fuzzy models synthesis and big data processing / A. Oliinyk, S. Skrupsky, S. Subbotin, O. Blagodariov, Ye. Gofman // Radio Electronics, Computer Science, Control. – 2016. – № 4. – P. 61–69. DOI: 10.15588/1607-3274-2016-4-8.
15. The model for estimation of computer system used resources while extracting production rules based on parallel computations / [A. A. Oliinyk, S. Yu. Skrupsky, V. V. Shkarupylo, S. A. Subbotin] / / Radio Electronics, Computer Science, Control. – 2017. – № 1. – С. 142–152. DOI: 10.15588/1607-3274-2017-1-16.
16. Oliinyk A. Parallel Computer System Resource Planning for Synthesis of Neuro-Fuzzy Networks / A. Oliinyk, S. Skrupsky, S. Subbotin // Recent Advances in Systems, Control and Information Technology. Advances in Intelligent Systems and Computing. – 2017. – Vol. 543. – P. 88–96. DOI: 10.1007/978-3-319-48923-0_12.
17. David Kirk Programming Massively Parallel Processors 3rd Edition. A Hands-on Approach / Kirk David, Hwu Wen-mei, 2016. – 576 p. ISBN: 9780128119860.
18. A comparison of approaches to large-scale data analysis / [Andrew Pavlo, Erik Paulson, Alexander Rasint et all] // International Conference on Management of Data. – 2009. – P. 165–178. DOI: 10.1145/1559845.1559865
19. Luiz Andre Barroso The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines / Luiz Andre Barroso, Urs Hoelzle // Synthesis Lectures on Computer Architecture. – 2009. – Volume 4, Issue 1. – P. 154. DOI: 10.2200/S00193ED1V01Y200905CAC006
20. Zaigham Mahmood Data Science and Big Data Computing: Frameworks and Methodologies / Zaigham Mahmood // Springer International Publishing. – 2016. – P. 332. DOI: 10.1007/978-3-319-31861-5
21. Salfner F. A survey of online failure prediction methods / F. Salfner, M. Lenk, M. Malek // ACM computing surveys. – 2010. – Vol. 42, Issue 3. – P. 1–42. DOI: 10.1145/1670679.1670680.
22. Shin Y. C. Intelligent systems : modeling, optimization, and control / C. Y. Shin, C. Xu. – Boca Raton : CRC Press, 2009. – 456 p. DOI: 10.1201/9781420051773.
23. Bishop C. M. Pattern recognition and machine learning / C. M. Bishop. – New York : Springer, 2006. – 738 p.
24. Gen M. Genetic algorithms and engineering design / M. Gen, R. Cheng. – New Jersey : John Wiley & Sons, 1997. – 352 p. DOI: 10.1002/9780470172254.
25. Yu X. Introduction to Evolutionary Algorithms (Decision Engineering) / X. Yu, M. Gen. – London : Springer, 2010. – 418 p. DOI: 10.1007/978-1-84996-129-5.
26. Ayala H. V. Cascaded evolutionary algorithm for nonlinear system identification based on correlation functions and radial basis functions neural networks / H. V. Ayala, L. D. Coelho // Mechanical Systems and Signal Processing. – 2016. – Vol. 68, Issue 6. – P. 376–378. DOI: 10.1016/j.ymssp.2015.05.022.
27. Abraham A. Swarm intelligence in data mining / A. Abraham, G. Grosan. – Berlin : Springer, 2006. – 267 p. DOI: 10.1007/978-3-540-34956-3.
28. Haroon Shakirat Oluwatosin Client-Server Model / Haroon Shakirat Oluwatosin // IOSR Journal of Computer Engineering. – 2014. – Volume 16, Issue 1. – P. 67–71. DOI: 10.9790/0661-16195771
29. Taylor R.N. Software Design and Architecture. The once and future focus of software engineering / R.N. Taylor, A. van der Hoek // International Conference on Software Engineering, 2007. – P. 226–243. DOI: 10.1109/FOSE.2007.21
30. An object-oriented architecture for extensible structural design software / [Rory Clune, Jerome J. Connor, John A. Ochsendorf, Denis Kelliher] // Computers & Structures. – 2012. – P. 1–17. DOI: 10.1016/j.compstruc.2012.02.002
31. Bernd Bruegge Object-Oriented Software Engineering Using UML, Patterns, and Java (3rd Edition) / Bernd Bruegge, Allen H. Dutoit. – Pearson, 2009. – 800 p. ISBN: 978-0136061250.
32. Qing Li Modeling and Analysis of Enterprise and Information Systems: From Requirements to Realization / Qing Li, Yu-Liu Chen. – Springer Berlin Heidelberg, 2009. – 405 p. DOI: 10.1007/978-3-540-89556-5
33. Hassan Gomaa Designing Software Product Lines with UML 2.0: From Use Cases to Pattern-Based Software Architectures / Hassan Gomaa. – Springer, 2006. – 440 p. DOI: 10.1109/SPLINE.2006.1691600
34. Making database systems usable / [H. V. Jagadish, Adriane Chapman, Aaron Elkiss, et al] // International Conference on Management of Data. – 2007. – P. 13–24. DOI: 10.1145/1247480.1247483
35. Sumathi S. Fundamentals of Relational Database Management Systems / S. Sumathi, S. Esakkirajan. – Springer-Verlag Berlin Heidelberg, 2007. – 792 p. DOI: 10.1007/978-3-540-48399-1
36. Lu Qin Keyword search in databases: the power of RDBMS / Lu Qin, Jeffrey Xu Yu, Lijun Chang // International Conference on Management of Data. – 2009. – P. 681–693. DOI: 10.1145/1559845.1559917
37. Subbotin S. Individual prediction of the hypertensive patient condition based on computational intelligence / S. Subbotin, A. Oliinyk, S. Skrupsky // Information and Digital Technologies : International Conference IDT'2015, Zilina, 7–9 July 2015 : proceedings of the conference. – Zilina : Institute of Electrical and Electronics Engineers, 2015. – P. 336–344. DOI: 10.1109/DT.2015.7222996.
38. Oliinyk A. O. Experimental Investigation with Analyzing the Training Method Complexity of Neuro-Fuzzy Networks Based on Parallel Random Search / A. O. Oliinyk, S. Yu. Skrupsky, S. A. Subbotin // Automatic Control and Computer Sciences. – 2015. – Vol. 49, Issue 1. – P. 11–20. DOI: 10.3103/S0146411615010071.
39. Intelligent information technology of automated diagnostic and pattern recognition systems development: Monograph / [S. A Subbotin, A. Oliinyk, Ye. Gofman et all]. – Kharkov : "Company Smith", 2012. – 317 p. (In russian).
40. Smith S. Genetic and Evolutionary Computation: Medical Applications / S. Smith, S. Cagnoni. – Chicehster : John Wiley & Sons, 2011. – 250 p. DOI: 10.1002/9780470973134.
41. Michael Creel Multi-core CPUs, Clusters, and Grid Computing: A Tutorial / Michael Creel, William L. Goffe // Computational Economics, 2008. – Volume 32, Issue 4. – P. 353–382. DOI: 10.1007/s10614-008-9143-5
42. Kshitij Gupta A study of Persistent Threads style GPU programming for GPGPU workloads / Kshitij Gupta, Jeff A. Stuart, John D. Owens // Parallel Computing. – 2012. – P. 1–14. DOI: 10.1109/InPar.2012.6339596
43. NVIDIA CUDA Compute Unified Device Architecture 5.5. Santa Clara : NVIDIA Corporation, 2014. – 117 p.

Article was submitted 28.08.2017.
After revision 15.10.2017.

Олійник А. О.[1], Субботін С. О.[2], Скрупський С. Ю.[3], Льовкін В. М.[4], Зайко Т. А.[5]

[1]Канд.техн.наук, доцент кафедри програмних засобів, Запорізький національний технічний університет, Запоріжжя, Україна

[2]Д-р техн. наук, завідувач кафедри програних засобів, Запорізький національний технічний університет, Запоріжжя, Україна

[3]Канд. техн. наук, доцент кафедри комп'ютерних систем та мереж, Запорізький національний технічний університет, Запоріжжя, Україна

[4]Канд. техн. наук, доцент кафедри програних засобів, Запорізький національний технічний університет, Запоріжжя, Україна

[5]Канд. техн.наук, ст. викладач кафедри програних засобів, Запорізький національний технічний університет, Запоріжжя, Україна

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ СИНТЕЗУ ДІАГНОСТИЧНИХ МОДЕЛЕЙ НА ОСНОВІ ПАРАЛЕЛЬНИХ ОБЧИСЛЕНЬ

Актуальність. Вирішено задачу автоматизації процесу синтезу діагностичних моделей при обробці великих масивів даних на основі паралельних обчислень. Об'єкт дослідження – процес синтезу діагностичних моделей. Предмет дослідження – методи та інформаційні технології синтезу діагностичних моделей.

Мета роботи полягає в створенні інформаційної технології синтезу діагностичних моделей.

Метод. Запропоновано інформаційну технологію синтезу діагностичних моделей, що представляє собою сукупність діаграм, які описують у графічному вигляді структурні елементи системи, а також поведінкові аспекти їх взаємодії на різних етапах побудови моделей об'єктів діагностування. Запропонована інформаційна технологія дозволяє виконувати побудову розподілених систем діагностування, в яких обчислювально складні етапи синтезу діагностичних моделей виконуються на високопродуктивному серверному обладнанні, що дозволяє істотно підвищити практичний поріг використання систем діагностування при обробці великих масивів даних, здатних вирішувати завдання редукції даних навчальної вибірки, видобування правил, побудови і донавчання діагностичних моделей.

Результати. Розроблено програмне забезпечення, яке реалізує запропоновану інформаційну технологію і дозволяє синтезувати діагностичні моделі на основі заданих наборів даних.

Висновки. Проведені експерименти підтвердили працездатність запропонованої інформаційної технології і дозволяють рекомендувати її для використання на практиці при обробці великих масивів даних для технічного і біомедичного діагностування. Перспективи подальших досліджень можуть полягати в модифікації розробленої технології шляхом впровадження в неї інших методів синтезу діагностичних моделей.

Ключові слова: вибірка даних, діагностування, видобування правил, відбір ознак, паралельні обчислення, синтез моделей.

Олейник А. А.[1], Субботин С. А.[2], Скрупский С. Ю.[3], Левкин В. Н.[4], Зайко Т. А.[5]

[1]Канд.техн.наук, доцент кафедры программных средств, Запорожский национальный технический университет, Запорожье, Украина

[2]Д-р техн. наук, заведующий кафедрой программных средств, Запорожский национальный технический университет, Запорожье, Украина

[3]Канд. техн. наук, доцент кафедры компьютерных систем и сетей, Запорожский национальный технический университет, Запорожье, Украина

[4]Канд. техн. наук, доцент кафедры программных средств, Запорожский национальный технический университет, Запорожье, Украина

[5]Канд. техн. наук, ст. преподаватель кафедры программных средств, Запорожский национальный технический университет, Запорожье, Украина

ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ СИНТЕЗА ДИАГНОСТИЧЕСКИХ МОДЕЛЕЙ НА ОСНОВЕ ПАРАЛЛЕЛЬНЫХ ВЫЧИСЛЕНИЙ

Актуальность. Решена задача автоматизации процесса синтеза диагностических моделей при обработке больших массивов данных на основе параллельных вычислений. Объект исследования – процесс синтеза диагностических моделей. Предмет исследования – методы и информационные технологии синтеза диагностических моделей.

Цель работы заключается в создании информационной технологии синтеза диагностических моделей.

Метод. Предложена информационная технология синтеза диагностических моделей, представляющая собой совокупность диаграмм, описывающих в графическом виде структурные элементы системы, а также поведенческие аспекты их взаимодействия на различных этапах построения моделей объектов диагностирования. Предложенная информационная технология позволяет выполнять построение распределенных систем диагностирования, в которых вычислительно сложные этапы синтеза диагностических моделей выполняются на высокопродуктивном серверном оборудовании, что позволяет существенно повысить практический порог использования систем диагностирования при обработке больших массивов данных, способных решать задачи редукции данных обучающей выборки, извлечения правил, построения и дообучения диагностических моделей.

Результаты. Разработано программное обеспечение, которое реализует предложенную информационную технологию и позволяет синтезировать диагностические модели на основе заданных наборов данных.

Выводы. Проведенные эксперименты подтвердили работоспособность предложенной информационной технологии и позволяют рекомендовать ее для использования на практике при обработке больших массивов данных для технического и биомедицинского диагностирования. Перспективы дальнейших исследований могут заключаться в модификации разработанной технологии путем внедрения в нее других методов синтеза диагностических моделей.

Ключевые слова: выборка данных, диагностирование, извлечение правил, отбор признаков, параллельные вычисления, синтез моделей.

REFERENCES

1. Price C. Computer based diagnostic systems. London, Springer, 1999, 136 p. DOI: 10.1007/978-1-4471-0535-0.
2. Bow S. Pattern recognition and image preprocessing. New York, Marcel Dekker Inc., 2002, 698 p. DOI: 10.1201/9780203903896.
3. Sammut C., Webb G. I. eds. Encyclopedia of machine learning. New York, Springer, 2011, 1031 p. DOI: 10.1007/978-0-387-30164-8.
4. Bezdek J. C. Pattern Recognition with Fuzzy Objective Function Algorithms. N.Y., Plenum Press, 1981, 272 p. DOI: 10.1007/978-1-4757-0450-1.
5. Sobhani-Tehrani E., Khorasani K. Fault diagnosis of nonlinear systems using a hybrid approach. New York, Springer, 2009, 265 p. (Lecture notes in control and information sciences ; № 383). DOI: 10.1007/978-0-387-92907-1.
6. Bodyanskiy Ye., Vynokurova O. Hybrid adaptive wavelet-neuro-fuzzy system for chaotic time series identification, Information Sciences, 2013, Vol. 220, pp. 170–179. DOI: 10.1016/j.ins.2012.07.044.
7. Subbotin S., Oliinyk A. The Dimensionality Reduction Methods Based on Computational Intelligence in Problems of Object Classification and Diagnosis, *Recent Advances in Systems, Control*

and Information Technology. Advances in Intelligent Systems and Computing, 2017, Vol. 543, pp. 11–19. DOI: 10.1007/978-3-319-48923-0_2.
8. Subbotin S., Oliinyk A. The Sample and Instance Selection for Data Dimensionality Reduction, *Recent Advances in Systems, Control and Information Technology. Advances in Intelligent Systems and Computing*, 2017, Vol. 543, P. 97–103. DOI: 10.1007/978-3-319-48923-0_13.
9. Oliinyk A. A., Skrupsky S. Yu., Shkarupylo V. V., Blagodariov O. Parallel multiagent method of big data reduction for pattern recognition, *Radio Electronics, Computer Science, Control*, 2017, No. 2, pp. 82–92.
10. Oliinyk A. Production rules extraction based on negative selection, *Radio Electronics, Computer Science, Control*, 2016, Vol. 1, pp. 40–49. DOI: 10.15588/1607-3274-2016-1-5.
11. Oliinyk A., Subbotin S. A. The decision tree construction based on a stochastic search for the neuro-fuzzy network synthesis, *Optical Memory and Neural Networks (Information Optics)*, 2015, Vol. 24, No. 1, pp. 18–27. DOI: 10.3103/S1060992X15010038.
12. Oliinyk A., Subbotin S. A. A stochastic approach for association rule extraction, *Pattern Recognition and Image Analysis*, 2016, Vol. 26, No. 2, pp. 419–426. DOI: 10.1134/S1054661816020139.
13. Oliinyk A. O., Skrupsky S. Yu., Subbotin S. A. Using Parallel Random Search to Train Fuzzy Neural Networks, *Automatic Control and Computer Sciences*, 2014, Vol. 48, Issue 6, pp. 313–323. DOI: 10.3103/S0146411614060078.
14. Oliinyk A., Skrupsky S., Subbotin S., Blagodariov O., Gofman Ye. Parallel computing system resources planning for neuro-fuzzy models synthesis and big data processing, *Radio Electronics, Computer Science, Control*, 2016, Vol. 4, pp. 61–69. DOI: 10.15588/1607-3274-2016-4-8.
15. Oliinyk A. A. Skrupsky S. Yu., Shkarupylo V. V., Subbotin S. A. The model for estimation of computer system used resources while extracting production rules based on parallel computations, *Radio Electronics, Computer Science, Control*, 2017, No. 1, pp. 142–152. DOI: 10.15588/1607-3274-2017-1-16.
16. Oliinyk A., Skrupsky S., Subbotin S. Parallel Computer System Resource Planning for Synthesis of Neuro-Fuzzy Networks, *Recent Advances in Systems, Control and Information Technology. Advances in Intelligent Systems and Computing*, 2017, Vol. 543, pp. 88–96. DOI: 10.1007/978-3-319-48923-0_12.
17. David Kirk, Hwu Wen-mei Programming Massively Parallel Processors 3rd Edition. A Hands-on Approach, 2016, 576 p. ISBN: 9780128119860.
18. Andrew Pavlo, Erik Paulson, Alexander Rasin, Daniel J. Abadi, David J. DeWitt A comparison of approaches to large-scale data analysis, *International Conference on Management of Data*, 2009, pp. 165–178. DOI: 10.1145/1559845.1559865
19. Luiz Andre Barroso, Urs Hoelzle The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines, *Synthesis Lectures on Computer Architecture*, 2009, Volume 4, Issue 1, pp. 154. DOI: 10.2200/S00193ED1V01Y200905CAC006
20. Zaigham Mahmood Data Science and Big Data Computing: Frameworks and Methodologies, *Springer International Publishing*, 2016, P. 332. DOI: 10.1007/978-3-319-31861-5
21. Salfner F., Lenk M., Malek M. A survey of online failure prediction methods, *ACM computing*, 2010, Vol. 42, Issue 3, pp. 1–42. DOI: 10.1145/1670679.1670680.
22. Shin Y. C., Xu C. Intelligent systems : modeling, optimization, and control. Boca Raton, CRC Press, 2009, 456 p. DOI: 10.1201/9781420051773.
23. Bishop C. M. Pattern recognition and machine learning. New York, Springer, 2006, 738 p.
24. Gen M., Cheng R. Genetic algorithms and engineering design. New Jersey, John Wiley & Sons, 1997, 352 p. DOI: 10.1002/9780470172254.
25. Yu X., Gen M. Introduction to Evolutionary Algorithms (Decision Engineering). London, Springer, 2010, 418 p. DOI: 10.1007/978-1-84996-129-5.
26. Ayala H. V., Coelho L. D. Cascaded evolutionary algorithm for nonlinear system identification based on correlation functions and radial basis functions neural networks, *Mechanical Systems and Signal Processing*, 2016, Vol. 68, Issue 6, pp. 376–378. DOI: 10.1016/j.ymssp.2015.05.022.
27. Abraham A., Grosan G. Swarm intelligence in data mining. Berlin, Springer, 2006, 267 p. DOI: 10.1007/978-3-540-34956-3.
28. Haroon Shakirat Oluwatosin Client-Server Model, *IOSR Journal of Computer Engineering*, 2014, Volume 16, Issue 1, pp. 67–71. DOI: 10.9790/0661-16195771
29. Taylor R. N., A. van der Hoek Software Design and Architecture. The once and future focus of software engineering, *International Conference on Software Engineering*, 2007, pp. 226–243. DOI: 10.1109/FOSE.2007.21
30. Rory Clune, Jerome J. Connor, John A. Ochsendorf, Denis Kelliher An object-oriented architecture for extensible structural design software, *Computers & Structures*, 2012, pp. 1–17. DOI: 10.1016/j.compstruc.2012.02.002
31. Bernd Bruegge, Allen H. Dutoit. Object-Oriented Software Engineering Using UML, Patterns, and Java (3rd Edition). Pearson, 2009, 800 p. ISBN: 978-0136061250.
32. Qing Li, Chen Yu-Liu Modeling and Analysis of Enterprise and Information Systems: From Requirements to Realization. Springer Berlin Heidelberg, 2009, 405 p. DOI: 10.1007/978-3-540-89556-5
33. Hassan Gomaa Designing Software Product Lines with UML 2.0: From Use Cases to Pattern-Based Software Architectures. Springer, 2006, 440 p. DOI: DOI: 10.1109/SPLINE.2006.1691600
34. Jagadish H. V., Adriane Chapman, Aaron Elkiss, Magesh Jayapandian, Yunyao Li Making database systems usable, *International Conference on Management of Data*, 2007, pp. 13–24. DOI: 10.1145/1247480.1247483
35. Sumathi S., Esakkirajan S. Fundamentals of Relational Database Management Systems. Springer-Verlag Berlin Heidelberg, 2007, 792 p. DOI: 10.1007/978-3-540-48399-1
36. Lu Qin, Jeffrey Xu Yu, Lijun Chang Keyword search in databases: the power of RDBMS, *International Conference on Management of Data*, 2009, pp. 681–693. DOI: 10.1145/1559845.1559917
37. Subbotin S., Oliinyk A., Skrupsky S. Individual prediction of the hypertensive patient condition based on computational intelligence, *Information and Digital Technologies : International Conference IDT'2015, Zilina, 7–9 July 2015 : proceedings of the conference*. Zilina, Institute of Electrical and Electronics Engineers, 2015, pp. 336–344. DOI: 10.1109/DT.2015.7222996.
38. Oliinyk A. O., Skrupsky S. Yu., Subbotin S. A. Experimental Investigation with Analyzing the Training Method Complexity of Neuro-Fuzzy Networks Based on Parallel Random Search, *Automatic Control and Computer Sciences*, 2015, Vol. 49, Issue 1, pp. 11–20. DOI: 10.3103/S0146411615010071.
39. Subbotin S. A., Oliinyk A., Gofman Ye., Zaitsev S., Oliinyk O. Intelligent information technology of automated diagnostic and pattern recognition systems development : Monograph. Kharkov, "Company Smith", 2012, 317 p. (In russian).
40. Smith S., Cagnoni S. Genetic and Evolutionary Computation: Medical Applications. Chicehster, John Wiley & Sons, 2011, 250 p. DOI: 10.1002/9780470973134.
41. Michael Creel, William L. Goffe Multi-core CPUs, Clusters, and Grid Computing: A Tutorial, *Computational Economics*, 2008, Volume 32, Issue 4, pp. 353–382. DOI: 10.1007/s10614-008-9143-5
42. Kshitij Gupta, Jeff A. Stuart, John D. Owens A study of Persistent Threads style GPU programming for GPGPU workloads, *Parallel Computing*, 2012, pp. 1–14. DOI: 10.1109/InPar.2012.6339596
43. NVIDIA CUDA Compute Unified Device Architecture 5.5. Santa Clara, NVIDIA Corporation, 2014, 117 p.