УДК 004.8

Т.І. Завалій, Ю.В. Нікольський, Т.В. Шестакевич
Інститут комп'ютерних наук та інформаційних технологій,
кафедра інформаційних систем та мереж

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ РЕЗУЛЬТАТІВ ПСИХОЛОГІЧНОГО ТЕСТУВАННЯ

Результати психологічного тестування, які є предметом дослідження, використовують для прийняття рішень щодо допуску фахівців до роботи операторами енергетичних мереж. Дані такого тестування містять багато різних показників, які перебувають між собою у складній залежності. Наведено результати дослідження, здійсненого з метою встановлення загальних закономірностей, які потрібно враховувати у разі прийняття рішення про допуск.

The psychological testing results, which were analized by the authors, are used for a decision-making about admitting the power engineering specialists to work. Such testing data contains a lot of different attributes with dependencies beetween them. The authors present the results of research conducted for the purpose of mining general patterns, which should be taken into account in the case of decision-making about admittance.

Постановка проблеми у загальному вигляді

Наведено результати аналізу даних психологічного тестування групи осіб, що працюють операторами енергетичних мереж. Таке тестування здійснюють раз на чотири роки для оцінювання психофізіологічних та особистісних якостей працівника. Результати тестів опрацьовує психолог, який за певною шкалою оцінює характерні особливості працівника – основні потреби, нахили, здібності, темперамент, мотиви його діяльності тощо. Психолог визначає професійну придатність фахівця. Крім того, професійні якості працівника – успішність та надійність – оцінюють два його керівники, відповідаючи на запитання спеціальної анкети.

Основними недоліками даних психологічного тестування є відсутність частини даних, що стосується окремих працівників, а також суб'єктивність оцінок, даних керівниками на основі власного досвіду. Накопичені в анкетах дані різнотипні і перед здійсненням аналізу потребують додаткового опрацювання.

Метою здійснених досліджень даних психологічного тестування є виявлення характерних особливостей працівника, які впливають на надійність його роботи та успішність. За результатати дослідження даних, наведеними у цій статті, можна вдосконалити процес тестування та виявити такі професійно важливі якості фахівця, яким належить приділяти увагу як під час тестувань, так і у разі прийняття кадрових рішень.

Для досягнення мети дослідження необхідно відібрати та попередньо опрацювати дані, проаналізувати дані з метою пошуку прихованих залежностей та оформити їх, оцінити ці залежності. Для інтелектуального аналізу даних пропонується використати технологію наближених множин.

Аналіз останніх досліджень

Загальна схема інтелектуального аналізу даних та проблема відсутніх даних

Інтелектуальний аналіз даних (DM, data mining) є складовою частиною процесу *видобування знань з баз даних* (KDD, knowledge discovery in data bases). Він дає змогу розкрити суть прихованих залежностей в даних, виявити взаємні впливи між властивостями об'єктів, інформація про які зберігається в базах даних, виділити закономірності, властиві певному набору даних.

Такий аналіз, завдяки використанню спеціальних методів машинного навчання та математичної статистики – від індукції правил і нейронних мереж до регресійного та дискримінантного аналізу – дає широкий спектр можливостей дослідження даних. Результатом процесу видобування знань вважають нові нетривіальні залежності, властивості даних, які можна інтерпретувати та використати на практиці. У загальному випадку результати видобування знань подають шаблонами, вирішувальними чи асоціативними правилами, функціональними залежностями.

Актуальність проблеми дослідження та опрацювання даних підтверджується широким практичним та комерційним використанням систем інтелектуального аналізу. Найчастіше їх застосовують у науковій сфері та бізнесі.

Такі системи дають можливість аналізувати поведінку клієнтів, досліджувати поділ клієнтів на групи, будувати різноманітні прогнозуючі моделі, аналізувати ризики, виявляти махінації в банківській та страхувальній сферах. У науці системи інтелектуального аналізу даних використовують для оцінювання результатів досліджень, аналізу масивів експериментальних даних, побудови різноманітних моделей тощо. Найвідоміші застосування таких систем у медицині, молекулярній генетиці та генній інженерії, прикладній хімії, фізиці.

У загальному випадку процес видобування знань може бути ітеративним і складається з чотирьох основних кроків.

1. Відбирання даних.
2. Попереднє опрацювання даних.
3. Інтелектуальний аналіз даних.
4. Оцінювання та інтерпретація побудованих моделей та знайдених залежностей.

Послідовність етапів [1] видобування знань зображено на рис.1.

Рис. 1. Загальна схема процесу видобування знань

На етапі відбирання даних для розв'язування певної задачі необхідно врахувати якість відібраних даних. Похибки, які виникають під час числового розв'язування задачі, поділяють на початкові, або неусувні похибки, похибки числового методу та обчислювальні похибки – похибки заокруглення та математичних дій. Повна похибка отриманого розв'язку залежить від усіх вказаних похибок.

У складних системах енергетики співвідношення між складовими похибки становить [2]: 2–3 % – через неточність числового методу, 82–84 % – неточність початкових даних, 14–15 % – неточність обчислень. Отже, якість початкових даних найбільше впливає на якість розв'язку.

Об'єкти предметної області описують множинами їхніх властивостей. Найзручніше подавати інформацію про властивості об'єктів таблицями, стовпці яких позначені іменами властивостей, а елементи рядків містять значення властивостей. Рядок таблиці у термінах машинного навчання є *прикладом,* який позначають через u, а множину усіх прикладів – U. Стовпці таблиці називають *атрибутами.* Позначимо їх через a, а множину всіх атрибутів – A. Таку таблицю ще називають *інформаційною системою*. Якщо у таблиці визначено *атрибут прийняття рішення* d, значення якого вказує на належність прикладу u до певного класу $d = d(u)$, то така таблиця є *таблицею прийняття рішень*. Тоді всі атрибути, крім атрибутів прийняття рішень, називають *умовними атрибутами*. Через $a(u)$ позначають значення атрибута a об'єкта u.

Проблема відсутніх даних у таблицях та невідомих значень атрибутів у прикладах з'являється тоді, коли існують приклади, визначені не на всій множині атрибутів, тобто хоча б одне $a(u)$ невідоме. Під невідомим розуміємо таке значення атрибута, визначити яке вже немає можливості, оскільки неможливо повторити умови, в яких були отримані всі інші дані у таблиці. Це значення може бути довизначене на основі певних міркувань, формуванню яких присвячені подальші дослідження.

Відсутність в таблиці значення атрибута a прикладу u позначатимемо як $a(u) = *$. Приклади, подані таблицями прийняття рішень, можуть мати невідомі значення як умовних атрибутів, так і атрибута прийняття рішення. Надалі розглядатимемо приклади, у яких можуть бути невідомими лише значення умовних атрибутів. Таблицю з даними називатимемо заповненою, якщо відомі значення всіх атрибутів усіх прикладів. Незаповнена таблиця матиме невідомим принаймні одне значення атрибута принаймні одного прикладу. Таблиця є порожньою, якщо вона не містить даних.

Відсутність даних, додатково до вже вказаних типів похибок, також впливає на результати досліджень, які здійснюють з метою побудови систем прийняття рішень. Відомо [3], що для статистичного аналізу до 1% відсутніх даних вважають тривіальним випадком, рівень у 1–5 % відсутніх даних вважають прийнятним для опрацювання, для даних з 5–15 % невідомими значеннями атрибутів необхідні складні методики дослідження. Найістотніше на результат аналізу даних може вплинути рівень відсутніх даних, який перевершує 15 % від їх загальної кількості.

Для інтелектуального аналізу таблиць даних, атрибути в яких мають невідомі значення, застосуємо порівняно новий підхід, який ґрунтується на понятті *наближеної множини* (rough set) [4]. Наближені множини – це символьна індуктивна методологія, яка поряд з нейронними мережами, розмитими множинами, генетичними алгоритмами належить до методологій *м'яких обчислень* (soft computing), які успішно застосовують в інтелектуальному аналізі даних та машинному навчанні.

Алгоритми, що використовують теорію наближених множин, застосовні для всього процесу видобування знань з таблиць даних. Зокрема, можна виділити алгоритми попереднього опрацювання даних – дискретизації, групування, доповнення невідомих значень тощо. На етапі інтелектуального аналізу застосовують алгоритми виведення набору *шаблонів* (patterns) для навчання без вчителя, чи набору правил у разі навчання з учителем. Математичний апарат теорії наближених множин забезпечує також механізми безпосередньої роботи з відсутніми даними.

Причини появи у таблицях невідомих значень атрибутів є такими [5,6].

1. Недбалість осіб, що збирають або вносять дані у таблиці, спричинена особистими рисами або відсутністю фінансової зацікавленості.

2. Зміна множини атрибутів у процесі збирання даних.

3. Надходження даних з різних джерел, у яких об'єкти описані різними множинами атрибутів.

4. Фізична відсутність даних. Наприклад, особа, яка не отримала водійських прав, не має запису про серію та номер посвідчення водія.

5. Логічна відсутність даних. Наприклад, керівник підприємства не можле вказати в анкеті прізвище свого начальника.

6. Помилки вимірювань та обмежені можливості апаратури.

7. Значення атрибута не належить допустимій множині його значень.

Існує ще одна причина появи у таблиці невідомих значень атрибутів. Щоб виконати вимогу конфіденційності та забезпечити анонімність, до даних вносять деякі спеціальні зміни. Ці зміни, що вносяться в процесі анонімізації, можуть призводити до зникнення частини даних. Анонімність даних актуальна в медичній сфері для збереження лікарської таємниці та уникнення ідентифікації пацієнтів за інформацією, яку аналізують. Анонімізацію у медицині необхідно виконувати перед відкриттям доступу до інформації з метою немедичного її використання, зокрема, аналізу даних.

Способи вирішення проблеми відсутніх даних

Приклади, які є описом об'єктів предметних областей та на основі яких доводиться приймати рішення, у своїй більшості містять невідомі значення атрибутів. У разі побудови систем прийняття рішень доводиться враховувати такі дані. Це пов'язано з тим, що здійснення додаткових замірів з метою отримання невідомих значень неможливе або вартісне. Тому можливі два шляхи використання неповних даних.

Перший з них – розроблення спеціальних методів, які врахують особливості зібраних даних, а другий – додаткове опрацювання даних та перетворення відповідних таблиць, що дає змогу застосувати універсальні методи побудови систем прийняття рішень.

Перший з цих підходів вимагає дослідження даних, здійснення тривалих експериментів, налаштування методу під дані, що є тривалим та незастосовним у реальних умовах. За другим підходом можна використати спеціальні методи опрацювання. Ці методи застосовні до широкого кола даних та використовують загальні підходи, не залежні від сфери застосування.

Виділяють [6] такі основні групи методів опрацювання таблиць із невідомими значеннями атрибутів:

- ігнорування відсутніх даних;
- видалення прикладів із невідомими значеннями атрибутів;
- доповнення відсутніх даних;
- безпосереднє опрацювання таблиць з відсутніми даними.

На рис. 2 показано методи опрацювання таблиць із відсутніми даними та вказано програмні продукти, у яких ці методи реалізовано (системи RSES 2.2 [7], LERS [8] та ROSETTA [9]).

Ігнорування відсутніх даних полягає у доповненні домена атрибута значенням, яким позначене невідоме значення цього атрибута.

Наприклад, нехай у таблиці з даними про особу є атрибут „освіта" з такими значеннями: 1 – „вища освіта", 2 – „середня спеціальна освіта". Якщо освіта певної особи невідома, то для позначення цього домен атрибута „освіта" можна поповнити нулем. Тоді доменом атрибута „освіта" стане множина $\{0,1,2\}$. Ігнорування невідомих значень дає змогу не змінювати розмірність таблиці прийняття рішень.

Видалення прикладів або атрибутів. Використовують два підходи до видалення даних. Перший з них здійснює спеціаліст, який, з огляду на свої знання та досвід, приймає рішення про видалення прикладу чи атрибуту. Такий підхід не є алгоритмічним, оскільки прийняття рішення про видалення залежить від досвіду людини-експерта. Другий підхід до видалення прикладів допустимий тоді, коли немає залежності між невідомими значеннями та причиною їх появи [3].

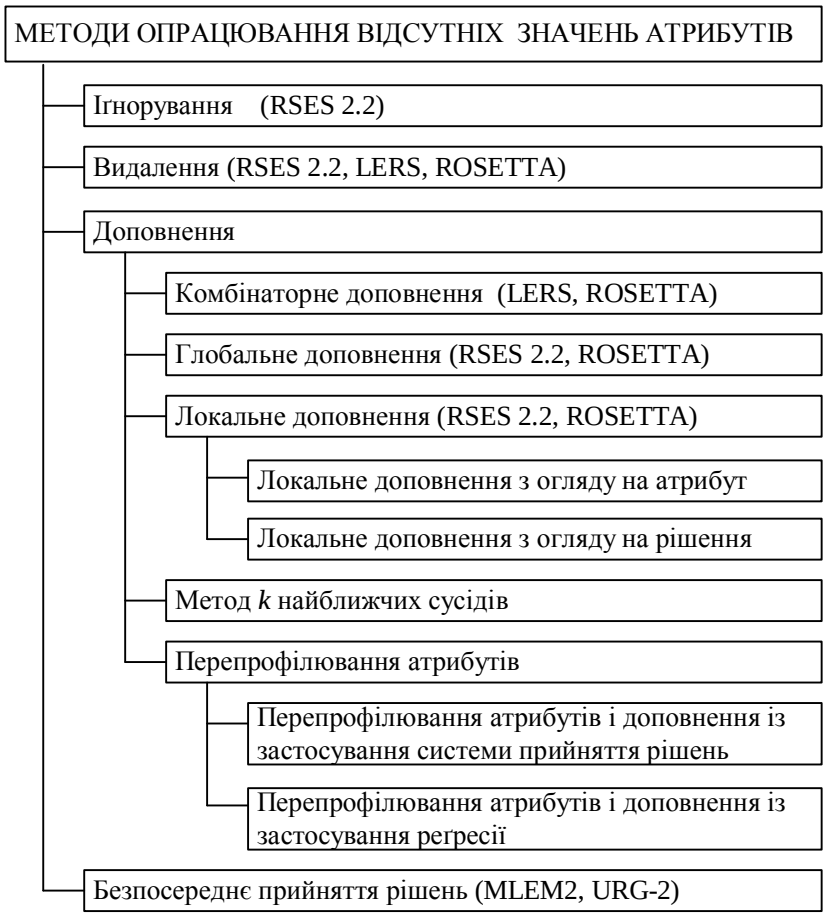

Рис. 2. Класифікація методів опрацювання відсутніх даних та програмні системи, у яких ці методи реалізовано

Проте, разом із прикладами з невідомими значеннями атрибутів можна вилучити й такі, що мають суттєві властивості такої таблиці прийняття рішень. Видалення даних може бути повним або частковим. У результаті повного видалення у таблиці будуть залишені лише заповнені рядки.

Пропорції та обсяги видалення прикладів із невідомими значеннями атрибутів залежать від конкретних даних та задач. У [3] описаний випадок видалення з таблиці понад 50% прикладів, у яких понід 30% значень атрибутів були невідомими. Вилучення прикладів чи атрибутів, що містять навіть одне невідоме значення, може значно зменшити розмірність таблиці.

Доповнення таблиць даних [6]. Невідомі значення атрибута заповнюють за певним критерієм, який формують на основі відомих його значень. У разі доповнення таблиці необхідно розрізняти дані, що об'єктивно існують, та такі, що не існують. До перших відносять такі дані, які можна отримати, але вони з певних причин не були внесені в таблицю. Наприклад, якщо невідомі дані про вік працівника, то їх можна доповнити на основі інших відомих даних.

Доповнення значень, які не існують, має два аспекти. Наприклад, у таблицю заносять дані про автомобілі певних осіб. Якщо у деякої особи немає автомобіля, то даних про колір машини, рік випуску та об'єм двигуна не існує. Отже, з одного боку, немає підстав для доповнення значень атрибутів, що повинні бути характеристиками автомобіля цієї особи. З іншого боку, аналіз даних вимагає заповненої таблиці.

Доповнення таблиць із відсутніми даними не змінює розмірності таблиці, але вносить інформаційний шум у дані. Універсальні методи доповнення таблиць з відсутніми даними дають змогу застосовувати відомі методи опрацювання заповнених таблиць прийняття рішень. Теоретичні

міркування, що стосуються заповнених таблиць прийняття рішень, можна застосувати для дослідження і незаповнених таблиць без аналізу невідомих значень.

Розглянемо такі методи доповнення таблиць з відсутніми даними (рис.2):

– комбінаторне доповнення;

– глобальне доповнення;

– локальне доповнення з огляду на атрибут та на рішення;

– доповнення методом k найближчих сусідів;

– перепрофілювання змінних і використання системи прийняття рішень.

Комбінаторне доповнення. Метод комбінаторного доповнення дає змогу доповнити таблицю заміною прикладу із невідомим значенням атрибута кількома прикладами із усіма відомими значеннями атрибутів. Обмеженням на застосування методу комбінаторного доповнення є велика кількість невідомих значень атрибутів і (або) велика потужність доменів атрибутів, значення яких невідомі.

Наприклад, нехай в таблиці присутній запис (41, *, Індія, добре), а доменом другого атрибута є множина {білий, жовтий, чорний}. Тоді цей приклад можна замінити такими новими прикладами: (41, білий, Індія, добре), (41, жовтий, Індія, добре), (41, чорний, Індія, добре). Подальші дії з опрацювання даних виконують над збільшеною кількістю прикладів.

Кількість додаткових прикладів, що утвориться із застосуванням методу комбінаторного доповнення, обчислюють за формулою

$$F = \sum_{i=1}^{n}\left(\prod_{j=1}^{m} z_{ij} - 1\right), \qquad (1)$$

де n – кількість прикладів у таблиці, m – кількість атрибутів таблиці, $z_{ij} = \begin{cases} 1, \text{ якщо } a_i(u_j) \neq *; \\ \left|V_{a_j}\right|, \text{якщо } a_i(u_j) = *, \end{cases}$ $a_i(u_j)$ – значення i-го прикладу на j-му атрибуті, $\left|V_{a_j}\right|$ – потужність домена j-го атрибута.

Таблиця 1

Приклад неповної таблиці

№	a_1	a_2	a_3	a_4
1	41	*	Індія	добре
2	41	*	Індія	*
3	22	жовтий	Європа	*

Для ілюстрації формули (1) розглянемо таблицю (табл. 1) з трьома прикладами та чотирма атрибутами a_1, a_2, a_3, a_4 потужностей 5, 3, 4 та 5, відповідно. Приклади цієї таблиці містять невідомі значення атрибутів. Обчислимо кількість додаткових прикладів, утворених за методом комбінаторного доповнення.

Оскільки $F = \sum_{i=1}^{3}\left(\prod_{j=1}^{4} z_{ij} - 1\right) = 20$, то табл. 1 після доповнення матиме 20 додаткових прикладів, а задана таблиця збільшить кількість прикладів з трьох до двадцяти трьох.

Метод комбінаторного доповнення застосовний лише для таблиць із відносно невеликою кількістю відсутніх даних і атрибутів з невеликими потужностями доменів.

Наприклад, нехай у таблиці із 194 прикладами та 38 атрибутами існують лише п’ять прикладів із невідомими значеннями атрибутів. У чотирьох прикладах невідомі значення одинадцяти атрибутів, чотири з яких складаються з чотирьох, а сім – з двох елементів. Ще в одного

прикладу невідомі значення семи атрибутів, кожний з яких має чотири елементи. Тоді кількість прикладів заданої таблиці збільшиться на $F = \sum_{i=1}^{194}\left(\prod_{j=1}^{38} z_{ij} - 1\right) = 147451$.

Глобальне доповнення. Таке доповнення використовують для заповнення відсутніх даних на основі відомих значень атрибутів. Для цього на основі усіх відомих значень атрибута обчислюють певний параметр s. Значенням параметра s для числових атрибутів може бути середнє або медіана, для символьних атрибутів – значення, що зустрічається найчастіше. Обчисленим параметром заміняють відсутні значення.

Метод глобального доповнення складається з таких кроків.

1. На основі відомих значень атрибута визначити параметр s.
2. Для кожного прикладу u такого, що $a(u) = *$, виконати $a(u) := s$.

Локальне доповнення з огляду на рішення. Розглянемо таблицю прийняття рішень. За методом глобального доповнення можна обчислювати параметр s локально, тобто на основі відомих значень лише того класу значень параметра прийняття рішень, до якого належить приклад.

Метод локального доповнення з огляду на рішення містить такі кроки.

1. Розбити множину прикладів U на такі підмножини $U_d = \{u \in U : d(u) = d\}$.
2. Для кожної підмножини обчислити значення параметра $s_d \in V_a$ на основі відомих значень атрибута.
3. У підмножині U_d для кожного прикладу u, для якого $a(u) = *$, виконати $a(u) := s_d$.

Головним недоліком методу локального доповнення з огляду на рішення є явище надмірного підсилення даних (overfitting), коли значний вплив на прийняте рішення отримують неістотні значення атрибутів.

Локальне доповнення з огляду на атрибут. В основу методу локального доповнення з огляду на рішення покладене припущення, що кожний умовний атрибут повинен корелювати з атрибутом прийняття рішення. Проте, такий зв'язок існує не завжди. Метод локального доповнення можна покращити, якщо замість атрибута прийняття рішення розглядати умовні атрибути із відсутніми значеннями. Оцінювання та врахування залежностей між атрибутами дає змогу уникнути надмірного підсилення даних.

Пошук пов'язаних між собою атрибутів ускладнений необхідністю оцінювати зв'язки між атрибутами не лише однакового, але й різного типу. Для оцінювання міри взаємозв'язку пари числових атрибутів можна використати коефіцієнт кореляції, а двох символьних атрибутів – ентропію. Проте немає ефективного методу порівняння між собою чисових та символьних атрибутів. Таке порівняння треба виконувати, якщо один із атрибутів таблиці містить числові дані, а всі інші атрибути – символьні. Може виявитись, що сильніші зв'язки існують між парами атрибутів різного типу, що можна не зауважити, якщо аналізувати взаємний вплив лише атрибутів однакового типу.

Доповнення за допомогою методу k найближчих сусідів. Підставою для застосування методу k найближчих сусідів є те, що приклади з близькими значеннями одних атрибутів найімовірніше мають близькі значення й на інших атрибутах. Розглянуті методи доповнення невідомих значень спирались на залежність між атрибутами у таблиці прийняття рішень, а подібність між прикладами майже не враховувалась.

Метод k найближчих сусідів ґрунтується на підході, першим кроком якого є підбір найближчих в певному сенсі прикладів. Для оцінювання подібності використовують усі дані про приклад. Для застосування методу k найближчих сусідів необхідно ввести поняття близькості прикладів. Якщо дані описано повністю, то припускають, що їх простір є метричним. Якщо таблиця

прийняття рішень містить дані з невідомими значеннями атрибутів, то на множині прикладів визначають функцію подібності, за допомогою якої визначають відстань від певного прикладу до усіх інших, а за її значенням обирають k найближчих прикладів.

Метод k найближчих сусідів для доповнення відсутніх значень складається з таких кроків [3].

1. Розбити множину прикладів U на дві підмножини U_m та U_c так, що підмножина U_m містить приклади з принаймні одним відсутнім значенням, а решта прикладів міститься в підмножині U_c.

2. Для кожного прикладу $u \in U_m$ виконати такі дії:

– поділити приклад на відомі u_c і невідомі u_m значення;

– знайти відстань між u_c та усіма прикладами з підмножини U_c;

– використати найближчі k сусідів для заповнення невідомих значень у прикладі u: невідомі значення атрибутів прикладу u замінити значенням параметра s, обчисленого на основі відомих значень атрибутів k найближчих сусідів. Такими параметрами можуть бути медіана, середнє або значення, що зустрічається найчастіше. У методі k найближчих сусідів можна керувати параметрами функції подібності об'єктів, величиною k та способами обчислення параметра s.

Порівняно із вже описаними методами опрацювання невідомих значень, метод k найближчих сусідів є найтривалішим за часом виконання. Його недоцільно застосовувати для опрацювання відносно великих за обсягом даних.

Перепрофілювання атрибутів і доповнення із застосуванням системи прийняття рішень. У цьому методі умовні атрибути з невідомими значеннями розглядають як атрибут прийняття рішень.

Метод перепрофілювання атрибутів і доповнення із застосуванням системи прийняття рішень складається з таких кроків.

1. Розбити множину атрибутів A на дві підмножини A_m та A_c так, що підмножина A_m містить атрибути з принаймні одним невідомим значенням, а решта атрибутів містяться в підмножині A_c.

2. Для кожного атрибута a ($a \in A_m$):

– розглядати атрибут a як атрибут прийняття рішень;

– розбити множину прикладів U на дві підмножини U_m та U_c так, що підмножина U_m містить приклади з принаймні одним невідомим значенням атрибуту, а решта прикладів містяться в підмножині U_c;

– для прийняття рішень вважати множину прикладів U_c навчальною та прийняти рішення щодо прикладів підмножини U_m.

Недолік методу виявляється тоді, коли потужність підмножини відомих атрибутів невелика порівняно із потужністю підмножини атрибутів із невідомими значеннями. Тобто застосування системи прийняття рішень на відносно малій навчальній множині робить результат практично непридатним для подальшого використання.

Перепрофілювання атрибутів і доповнення із застосуванням регресії. Існування зв'язків між невідомими та відомими значеннями можна використати для доповнення невідомих значень атрибута за допомогою регресійного аналізу [10]. У цьому випадку невідомі значення розглядають як залежні змінні, а методи регресійного аналізу використовують для їх визначення. Найпростішою у використанні є лінійна регресія. Якщо ж залежність між даними є нелінійною, застосування лінійної регресії може спотворити результат.

Для покращання результатів, які можна отримати застосуванням регресійного аналізу, можна використати множинну регресію [10,11], у якій обчислюють не один варіант визначення кожного невідомого значення, а множину можливих значень, утворюючи різні заповнені таблиці. Отримані таблиці опрацьовують методами для заповнених таблиць і вибирають найкращий результат.

Отже, доповнення невідомих значень є універсальним способом розв'язування задачі про неповний опис об'єктів. Водночас доповнення невідомих даних має небезпеку внесення істотних змін у дані, що ускладнить пошук зв'язків між умовними атрибутами та рішенням.

Безпосереднє прийняття рішень за допомогою неповних даних. Одним із способів безпосереднього опрацювання даних з невідомими значеннями є методи поділу, за допомогою яких таблицю прийняття рішень з відсутніми даними поділяють так, щоб утворити заповнені таблиці. Такі таблиці опрацьовують методами для заповнених таблиць.

Існують також методи опрацювання прикладів з невідомими значеннями атрибутів, які полягають у прийнятті рішень безпосередньо на основі неповних даних. Такий підхід реалізовано в алгоритмах MLEM2 [12], URG-2 [13]. Крім цього, деякі механізми оперування прикладами з неповним описом та прийняття рішень на їх основі забезпечує методологія наближених множин (див. розділ 2.4). Негативний аспект такого підходу полягає в індивідуальному налаштуванні алгоритму до заданого набору даних.

При опрацюванні неповних даних треба враховувати випадок, коли відсутність даних необхідна з метою забезпечення конфіденційності. У цьому випадку виконують анонімізацію даних.

Існує три основних підходи анонімізації таблиці прийняття рішення [14].

1. *Вилучення даних*. Такий підхід застосовують до таблиць, у яких є приклад, що істотно відрізняється від решти прикладів. Для уникнення ідентифікації відповідного об'єкта цей приклад вилучають з таблиці.

Наприклад, нехай серед певної групи осіб є небагато жінок і лише одна з них передпенсійного віку. Тоді усі інші дані цієї особи легко встановити. Також вилученню можуть підлягати атрибути. Будь-які вилучення даних зменшують розміри таблиці.

2. *Узагальнення даних*. За таким підходом зменшують потужність домена атрибута шляхом об'єднання кількох значень цього атрибута. Два приклади є нерозрізненними на атрибуті, якщо значення атрибута на різних прикладах збігаються або одне із них невідоме. Отже, два розрізненних до такого об'єднання приклади стають нерозрізненними за цим атрибутом.

Наприклад, нехай деяка множина осіб поділена за віком на групи. Якщо серед цих груп є дві найменш чисельні – „50–55 років" та „понад 55 років", то доцільно об'єднати їх в одну групу „понад 50 років". Тоді приклади, що містять дані про двох осіб, одна з яких пенсійного віку, стануть нерозрізненними за атрибутом „вік". Узагальнення не змінює розмірів таблиці.

3. *Замовчування даних*. У такий спосіб анонімність даних досягається забезпеченням умов нерозрізненності двох прикладів на атрибуті. Для цього прикладу u, який потрібно анонімізувати, ставлять у відповідність підмножину прикладів $U_s \in U$, які можна використати для ідентифікації u. У кожному прикладі з U_s видаляють значення відповідного атрибута.

Наприклад, якщо у групі осіб є небагато осіб із середньою освітою, то можна досягти нерозрізненності прикладів з таблиці за атрибутом „освіта". Замовчування дає змогу зберегти розмірність таблиці даних.

Технологія наближених множин

Порівняно нова методологія наближених множин природним чином пристосована до роботи з недосконалими даними. Теорія наближених множин створена Ж. Павлаком (Z. Pawlak) [11] як математичний інструмент для подолання суперечностей у даних і виявлення в них прихованих закономірностей або шаблонів.

Фундаментальний принцип алгоритму навчання на прикладах з використанням наближених множин полягає у виявленні надлишковості серед наявних ознак, що описують приклад. Так виявляють сильні ознаки, які впливають на класифікацію прикладу. Використанням нижнього та верхнього наближень цей алгоритм наближає певне поняття знизу та згори і в такий спосіб усуває різні суперечності та невизначеності. Під поняттям у цьому випадку розуміють клас, до якого належить множина прикладів.

Дані, що досліджують, подають у вигляді прикладів, зібраних у таблицю прийняття рішень. Нехай $U=\{u_1,u_2,\mathbf{K},u_n\}$ – непорожня скінченна множина прикладів, кожен з яких поданий рядком таблиці, $A=\{a_1,a_2,\mathbf{K},a_m\}$ – непорожня скінченна множина атрибутів та $a: U\to V_a$ для всіх $a\in A$. Множину V_a називають множиною значень, або доменом атрибута, а саму таблицю – *інформаційною системою.*

Інформаційну систему поповнюють атрибутом прийняття рішення d, за значенням якого приклади відносять до відповідного класу. Таблиця $\mathbf{A}=(U,A\cup\{d\})$ з класифікаційним атрибутом d є таблицею прийняття рішень. У загальному випадку така таблиця може мати більше ніж один класифікаційний атрибут.

На множині U прикладів таблиці визначають відношення нерозрізненності. Нехай $B\subseteq A$ – підмножина всієї множини атрибутів таблиці прийняття рішень $\mathbf{A}$, u та u' – приклади таблиці. Тоді $IND(B)=\{(u,u')\,|\,U\times U,\forall a\in B,a(u)=a(u')\}$ – відношення B-нерозрізненності. Якщо явно не задають множину B, то мають на увазі всю множину A атрибутів таблиці. Отже, приклади u та u' нерозрізненні, якщо однаковими є відповідні значення їхніх атрибутів. Відношення нерозрізненності симетричне, рефлексивне і транзитивне, а, отже, є відношенням еквівалентності. Класи еквівалентності, отримані на основі цього відношення, позначають через $[X]_B$, де $X\subseteq U$. Приклади цього класу еквівалентності нерозрізненні на множині атрибутів B.

Частина даних може бути надмірною або суперечливою, зокрема, якщо певні приклади еквівалентні на множині умовних атрибутів, але мають різні значення атрибута прийняття рішення. Ці приклади неможливо однозначно класифікувати. Кажуть, що вони належать *граничній області*. Такі суперечливі приклади вилучають з таблиці. Також вилучають приклади, для яких існує повністю, включно з атрибутом прийняття рішення, еквівалентний приклад.

Після цього вилучають стовпці, атрибути яких не впливають на класифікацію. Для аналізу залишають лише такі атрибути, від яких залежить класифікація прикладів таблиці. Множину атрибутів, що залишилися, називають *редуктом*. Іншими словами, редукт – це підмножина усіх атрибутів таблиці $\mathbf{A}$, які забезпечують той самий результат класифікації всіх прикладів таблиці, як і вся множина A. Атрибутам можна надавати ознаку важливості з інтервалу [0, 1].

З погляду якості кінцевого результату ефективним підходом є побудова так званих *динамічних* редуктів, які обчислюють для окремих частин таблиці. Для обчислення динамічних редуктів довільно формують k підтаблиць, і серед знайдених для них редуктів відбирають ті, які зустрічаються найчастіше. Цей підхід більшою мірою враховує особливості розподілу даних в таблиці.

Також визначають *об'єктно-залежний* редукт – мінімальний набір атрибутів, значення яких дають змогу відрізнити певний приклад з таблиці від решти прикладів. Знаходження редукта є NP-складною задачею. Однак існують достатньо ефективні методи, які дають змогу розв'язати цю задачу за прийнятний час. До таких методів належать, зокрема, *логічне виведення* (boolean reasoning) [14] та генетичний алгоритм [15].

Метод логічного виведення полягає у побудові *функції розрізнення* (discernibility function) та її спрощенні. У загальному випадку, функція розрізнення є такою булевою функцією m змінних

$g_A(a_1,...,a_m) = \wedge\{\vee c_{ij}^* \mid 1 \le j \le i \le n, c_{ij}^* \ne 0\}$, де $i, j = (1,2,\mathbf{K},n)$, m – кількість атрибутів таблиці $\mathbf{A}$, n – кількість прикладів, a_m – атрибут таблиці, c_{ij}^* – елемент спеціальної *матриці розрізнення* $M(A)$.

Матриця розрізнення є симетричною матрицею розмірів $n \times n$ з нульовою діагоналлю. Елемент c_{ij}^* є диз'юнкцією атрибутів, за значеннями яких відрізняються два приклади u_i та u_j з U. За методом логічного виведення ці приклади повинні мати різні значення атрибута прийняття рішення. Якщо приклади мають однакові значення атрибута прийняття рішення, то $c_{ij}^* = 0$.

Наступним кроком логічного виведення є зведення функції розрізнення до досконалої кон'юнктивної нормальної форми, атоми якої є атрибутами, що утворюють редукт. У випадку, коли це неможливо, функцію спрощують до кон'юнктивної нормальної форми мінімальної довжини та переводять у диз'юнктивну нормальну форму, диз'юнкти якої утворюють редукти.

Інший спосіб побудови редукта реалізовано в алгоритмі Джонсона [9]. Нехай R – множина атрибутів, які утворюють шуканий редукт, S – набір підмножин s_i атрибутів a_m за значеннями яких відрізняються два приклади u_i та u_j з U, $a_m \in A$, $s_i \subseteq A$. Ці підмножини атрибутів відповідають кон'юнктам функції розрізнення, яка будується при логічному виведенні. Тобто S є інакшим способом подання функції g_A. Для кожної підмножини s_i обчислюють коефіцієнт $w(s_i)$, який позначає вагу s_i в S. У [14] пропонується приймати вагу підмножини рівною кількості її входжень у S. Весь алгоритм пошуку редукта складається з п'яти кроків.

Крок 1. Покласти $R = \varnothing$.

Крок 2. Вибрати в S атрибут a, який максимізує значення $\sum w(s_i)$ всіх s_i, які містять a.

Крок 3. Додати a до R.

Крок 4. Видалити з S всі підмножини s_i, що містять a.

Крок 5. Якщо $S = \varnothing$, то R – шуканий редут, кінець. Інакше – перехід до кроку 2.

Якщо на кроці 5 припинити побудову редукта тоді, коли з набору S вилучено не всі елементи, то ми отримаємо так званий наближений редукт. Цей редукт характеризується *ступенем підтримки* (hitting fraction) – відсотком тих підмножин s_j від загальної кількості s_i, атрибути з яких увійшли до редукта R ($s_j \cap R \ne \varnothing$). Тобто, при побудові наближеного редукта враховуються не всі підмножини s_i, а лише заданий їх відсоток з більшою вагою w.

У реальних наборах даних навіть один приклад може вносити у дані шум і змінювати функцію нерозрізненності для таблиці, що, своєю чергою, відобразиться у редукті. Пошук наближеного редукта дає змогу зменшити вплив на знайдений редукт шумів і неточностей в даних. Такий редукт є „сильнішим", ніж звичайний та відображає загальніші закономірності у таблиці даних. Крім методу Джонсона для побудови наближеного редукта можна використати генетичний алгоритм зі спеціальною [9] функцією мети.

Наближені множини і відсутні значення

Відношення нерозрізненності є основним поняттям у теорії наближених множин. Його використовують для побудови спеціальних булевих функції, які використовують у процесі логічного виведення. У разі наявності прикладів з відсутніми значеннями атрибутів виникає

проблема порівняння таких прикладів із застосуванням відношення нерозрізненності. Для її вирішення теорія наближених множин пропонує деякі модифікації відношення нерозрізненності.

Першим з таких відношень є відношення толерантності, або симетричної подібності. Це відношення є натуральним розширенням відношення нерозрізненності. Відношення симетричної подібності застосовують як у методології наближених множин, так і в інших методах. Припускають, що відсутнє значення певного атрибута потенційно може бути будь-яким значенням з домену цього атрибута. Введемо поняття відношення толерантності згідно з [6]. Якщо $\mathbf{A} = (U, A)$ – інформаційна система та $B \subseteq A$, то відношення толерантності $TOL_{\mathbf{A}}(B) \subseteq U \times U$ на множині атрибутів B задають так

$$TOL_{\mathbf{A}}(B) = \{(u,u') \in U \times U \mid \forall a \in B:\ a(u) = a(u) \vee a(u') = * \vee a(u') = *\},$$

де символом $*$ позначене невідоме значення.

Використання відношення толерантності для порівняння прикладів з невідомими значеннями атрибутів є рівнозначним використанню методу комбінаторного доповнення (див. розділ 2.2) на етапі попереднього опрацювання даних. Перевага відношення толерантності полягає в тому, що не потрібно доповнювати таблицю новими прикладами, а достатньо застосувати відношення толерантності в процесі навчання чи аналізу.

Альтернативним відношенням є відношення несиметричної подібності, яке є точнішим за відношення симетричної подібності. Для задання несиметричної подібності вводять поняття прикладу-копії і прикладу-оригіналу. Наведемо визначення відношення несиметричної подібності [6]. Якщо $\mathbf{A} = (U, A)$ – інформаційна система та $B \subseteq A$, то відношення несиметричної подібності $SIM_{\mathbf{A}}(B) \subseteq U \times U$, утворене на множині атрибутів B, задають як $SIM_{\mathbf{A}}(B) = \{(u,u') \in U \times U \mid \forall a \in B: a(u) = a(u) \vee a(u') = *\}$.

Щоб приклад u був подібним до прикладу u', має виконуватись звичайна умова рівності значень відповідних атрибутів. Але необхідно, щоб приклад u' був „оригіналом" для прикладу u, тобто, був описаний принаймні на тих самих атрибутах, що й приклад u. Приклад u може мати більше атрибутів. Отже, він може мати більше невідомих значень атрибутів, ніж u'.

Введені відношення подібності трактують подібність прикладів у строго визначений спосіб. Вони не враховують такого важливого аспекту, як міра подібності прикладів між собою. Тому варто використовувати відношення розмитої подібності, яке дає змогу приписати прикладам з множини U міру подібності з проміжку [0, 1]. Це дає більші можливості для порівняння, ніж простий поділ прикладів на подібні та неподібні.

Найпоширенішою формою відношення розмитої подібності є ймовірнісна інтерпретація невідомих значень. Тобто вважають, що невідоме значення може з однаковою ймовірністю набувати будь-якого значення з домена атрибута. Відповідно до цього, для двох прикладів u та u' подібність значень атрибута a_i можна записати через коефіцієнт подібності

$$R_{a_i}(u,u') = \begin{cases} 1 & \text{, якщо } a(u) = a(u') \neq *, \\ 0 & \text{, якщо } a(u) \neq (a(u') \wedge a(u)) \neq (* \wedge a(u')) \neq *, \\ \dfrac{1}{|V_{a_i}|} & \text{, якщо } a(u) = (* \wedge a(u')) \neq (* \vee a(u)) \neq (* \wedge a(u')) = *, \\ \dfrac{1}{|V_{a_i}|^2} & \text{, якщо } a(u) = (* \wedge a(u')) = *. \end{cases}$$

Тоді відношення розмитої подібності $R_{\mathbf{A}}(B):U\times U\to[0,1]$ визначають на підмножині атрибутів B так: $R_{\mathbf{A}}(B)(u,u')=\prod_{a\in B}R_a(u,u')$ [6]. Так введене відношення подібності відповідає ймовірнісній інтерпретації невідомих значень як випадкових подій у класичному розумінні. Крім цього, з огляду на можливі суперечності, які можуть вноситись у процес виведення, покладають $R_{\mathbf{A}}(u,u)=1.$

З використанням вказаних відношень можна побудувати функції розрізнення на основі таблиць з неповними даними, а отже, здійснювати весь процес логічного виведення на таких даних.

Цілі статті

У статті наведено результати застосування методології наближених множин для аналізу неповних таблиць даних, отриманих в результаті психологічного тестування групи осіб. Задача здійсненого дослідження полягала в аналізі таблиці прийняття рішення відповідно до загального процесу видобування знань.

У результаті проведеного дослідження виконано такі основні завдання:

– виконано опис даних;

– здійснено попереднє опрацювання даних;

– знайдено атрибути, які найбільше впливають на прийняте рішення;

– виділено залежності в таблиці та побудувано правила;

– оцінено отримані результати.

Опис наявних даних здійснено з метою уточнення подальших завдань аналізу та для попереднього опрацювання даних. З усього набору даних сформовано єдину таблицю прийняття рішень. Дані трансформовано до формату, якого вимагає алгоритм аналізу.

У процесі попереднього опрацювання даних вилучено зайві дані, замінено тип частини даних, закодовано окремі значення, довизначено або вилучено невідомі значення. Пошук атрибутів, які найбільше впливають на прийняте рішення, та виділення залежностей в даних є ядром досліджень і здійснюється за допомогою технології наближених множин.

Наведено фраґмент побудованої множини правил прийняття рішень. Досліджено вплив на результати аналізу доповнення невідомих значень методом глобального доповнення. Для визначення впливовості атрибутів побудовано наближені редукти за методом Джонсона з різним значенням ступеня підтримки. Порівнянням результатів різних експериментів оцінено отримані результати, сформульовано висновки та надано рекомендації щодо подальших досліджень.

Основний матеріал

Підготовка даних для експериментів

Для аналізу обрано результати професійного психологічного тестування диспетчерського складу підприємства енергетичної галузі.

Разом опрацьовано та зібрано в таблицю дані тестування 206 осіб, з яких:

- чоловіків 184 (89.32%), жінок 22 (10.68%);
- диспетчерів оперативної диспетчерської служби 21 особа (10.19%),
- диспетчерів районних електромереж 166 осіб (80.58%),
- диспетчерів підстанцій 19 осіб (9.23%);
- здобули середню спеціальну освіту 100 осіб (48.54%), вищу – 106 осіб (51.46%).

За віком працівників розподілено на такі групи:

- від 20 до 29 років – 41 особа (19.9%);
- від 30 до 39 років – 68 осіб (33%);
- від 40 до 49 років – 51 особа (24.76%);

– від 50 до 59 років – 26 осіб (22.34%);
– більше 59 років – 0 осіб (0%).

Розподіл осіб по групах за стажем подано у табл. 2.

Таблиця 2

Розподіл працівників за стажем

Позначення	1	2	3	4	5	6	7
Стаж (років)	0-5	6-10	11-15	16-20	21-25	26-30	>30
Кількість осіб	110	36	30	14	10	6	0
%	53.39	17.48	14.56	6.8	4.85	2.92	0

Таблиця даних містить 206 прикладів та 65 атрибутів, тобто загальна кількість значень в таблиці із врахуванням невідомих значень становила 13390. Серед цих даних є 18 прикладів з невідомими значеннями атрибутів, що становить 8.7% від загальної кількості прикладів; невідомих значень атрибутів – 162, або 1.21%.

Структура набору даних зображена на рис. 3. Весь набір даних містить результати двох типів опитувань: оцінки психологічних якостей працівника та характеристики працівника як спеціаліста, виконаної його керівниками.

За своєю структурою множину умовних атрибутів можна розбити на такі підмножини.

1. Інформативні. Ці атрибути містять загальну інформацію про працівника.

2. Атрибути-харакеристики. Ці атрибути містять результати психологічних тестувань та оцінки цих результатів, а також інфомацію про характерні психологічні особливості працівника.

3. Проміжні оцінки. Ці атрибути містять узагальнюючі оцінки працівника, зроблені психологом на основі інформації з атрибутів-характеристик.

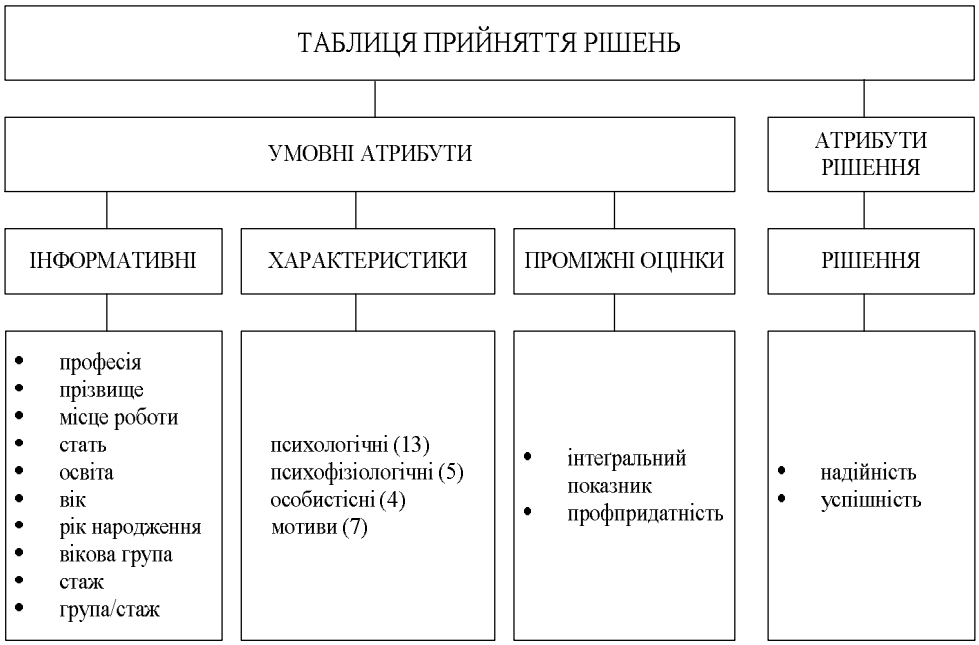

Рис. 3. Структура таблиці даних

Таблиця прийняття рішень з оцінкою психологічних якостей працівника містить атрибути, які поділено на чотири групи.

1. *Інформативні атрибути*. До інформативних віднесемо атрибути „професія”, „прізвище”, „місце роботи”, „стать”, „освіта”, „вік” „рік народження”, „вікова група”, „стаж”, „група/стаж”.

Атрибут „професія” може набувати таких значень: 1 (диспетчер центральної диспетчерської служби); 2 (диспетчер оперативної диспетчерської служби); 3 (начальник районної диспетчерської служби); 4 (старший диспетчер районної електромережі); 5 (диспетчер районної електромережі); 6 (диспетчер підстанції). Атрибут „місце роботи” набуває значення одного з чотирьох підрозділів магістральних електромереж (ЕМ) – ВЕМ, ПЕМ, ГЕМ, ОЕМ.

Атрибут „освіта” набуває значення 1 (вища) або 2 (середня спеціальна). Атрибут „вікова група” набуває значення однієї з таких вікових груп 1 (від 20 до 29 років), 2 (від 30 до 39 років), 3 (від 40 до 49 років), 4 (від 50 до 59 років), 5 (більше 59 років). Домен атрибута „група/стаж” утворений сімома групами за стажем з інтервалом у 5 років (див табл. 2).

2. *Атрибути-характеристики*. Ці атрибути містять інформацію про результати тестувань, пройдених працівниками. Всі атрибути-характеристики розбиті на такі підгрупи:

– психологічні характеристики: „концентрація уваги”, „переключення уваги”, „кількість помилок уваги”, „обсяг уваги”, „точність відтворення інформації”, „швидкість відтворення інформації”, „технічне мислення”, “швидкість технічного мислення”, „оперативна пам’ять”, “об’єм оперативної пам’яті”, „логічне мислення”, “якість логічного мислення”, „короткочасна пам’ять”;

– психофізіологічні характеристики: „швидкість реакції”, „стійкість реакції”, „рівень функціональних можливостей”, „сила нервової системи”, „рухливість нервових процесів”;

– особистісні характеристики: „потреба досягнень”, „активність”, „стиль керівництва”, „організаційні здібності”;

– мотиви: „мотив – гроші”, „мотив – соціальний статус”, „мотив – суспільна активність”, „мотив – суспільна корисність”, „мотив – творча активність”, „мотив – комфорт”, „мотив – спілкування”.

У таблиці даних для кожного з атрибутів-характеристик, крім мотивів та атрибута „стиль керівництва”, існує відповідний атрибут з оцінкою результату тесту за чотирибальною шкалою. Ці атрибути можуть набувати значення в (відмінно), д (добре), з (задовільно), н (незадовільно). Загалом до групи атрибутів-характеристик входять 50 атрибутів. Атрибут „стиль керівництва” теж набуває значення з домену [в, д, з, н]. Атрибути з інформацією про мотиви набувають значення 0, якщо такий мотив у працівника відсутній, значення 1 – у протилежному випадку.

3. *Проміжні оцінки*. До проміжних оцінок віднесено „інтегральний показник” (ІП) та „профпридатність”.

Атрибут „інтегральний показник” обчислено як середнє арифметичне значень атрибутів з результатами тестування психологічних та психофізіологічних якостей. При обчисленні цього значення не враховано атрибут „якість логічного мислення”. Значення інтегрального показника узагальнене значенням атрибута „оцінка ІП” за чотирибальною шкалою. Атрибут „оцінка ІП” набуває значення з домену [в, д, з, н].

Атрибут „профпридатність” оцінює психолог на основі інтегрального показника та особистісних характеристик працівника. Цей атрибут може набувати таких значень: бп (безумовно придатний), п (придатний), уп (умовно придатний), р (група ризику).

4. *Атрибути прийняття рішення*. Атрибутами прийняття рішень є „успішність” та „надійність”. Оцінка успішності та надійності роботи працівника здійснювалась його керівниками за п’ятибальною шкалою в процесі заповнення окремої анкети. Атрибути прийняття рішень набувають значень з домена [1, 2, 3, 4, 5].

Недоліки даних. Описані дані мають кілька недоліків.

По-перше, існує певна надлишковість даних. Більшість характеристик працівника подано в таблиці двома атрибутами. Один атрибут містить кількість балів, які є результатом відповідного тестування, інший – оцінку цього результату за певною шкалою. Тому загальна кількість умовних атрибутів складається з 63 атрибутів. На стадії відбирання даних необхідно вилучити атрибути, що містять отримані бали, та залишити для аналізу лише атрибути зі значеннями оцінок.

По-друге, у таблиці даних невідомі деякі значення атрибутів, через що незастосовні методи інтелектуального аналізу даних, які працюють лише із заповненими таблицями.

По-третє, оскільки тестування здійснювалось тільки раз на чотири роки, то це ускладнює відслідковування змін параметрів у часі. З огляду на вимоги конфіденційності, відсутність частини даних, а також необхідність уніфікації кодування значень атрибутів потрібно здійснити попереднє опрацювання даних.

Нарешті, загальним недоліком прийнятих рішень є їхня суб'єктивність. Власне, суб'єктивність прийнятих рішень вимагає розв'язування задачі знаходження залежностей у даних та атрибутів, які реально впливають на прийняття рішень. Знаходження таких атрибутів дасть змогу вдосконалити психологічні дослідження у майбутньому.

Детальніше розглянемо виконання етапів дослідження даних психологічних тестувань. Послідовність етапів дослідження даних відповідає загальному процесу видобування знань, зображеному на рис. 1. Кожний з етапів має свою специфіку, описану нижче.

Етап відбирання даних. Відбирання даних полягає у видаленні атрибутів таблиці, не важливих для аналізу. Перш за все, під час відбирання даних було вилучено атрибути „прізвище", „рік народження", „вік", „стаж" та „інтегральний показник". Вилучення атрибута „прізвище" забезпечує конфіденційність інформації. Атрибути „рік народження" та „вік" узагальнені атрибутом „вікова група", „стаж" – атрибутом „група/стаж", а „інтегральний показник" – атрибутом „оцінка ІП".

Також на етапі відбирання даних вилучено:

– 13 атрибутів, які містять результати психологічних тестів, оскільки ці атрибути узагальнені відповідними атрибутами з оцінками цих результатів;

– 5 атрибутів з результатами психофізіологічних тестів, оскільки вони узагальнені відповідними атрибутами з оцінками цих результатів;

– 3 атрибути з результатами тестування таких особистісних характеристик, як потреба досягнень, активність та організаційні здібності, оскільки вони узагальнені відповідними атрибутами з оцінками цих результатів.

Після відбирання даних загальна кількість атрибутів таблиці зменшена, і разом з атрибутом прийняття рішення становить 38 замість 65. Отже, отримана таблиця даних має 206 прикладів та 38 атрибутів.

Етап попереднього опрацювання даних. На цьому етапі було здійснено опрацювання атрибутів з оцінкою таких особистісних характеристик працівника, як „потреба досягнень", „активність", „стиль керівництва" та "організаційні здібності". Ці характеристики під час психологічних тестувань оцінювались за різними шкалами. Домени цих атрибутів перетворені з [1, 2, 3] на [в (відмінно), д (добре), з (задовільно)]. Такою заміною уніфіковано подання оцінок в цих атрибутах, оскільки були різними шкали, за якими ставилися ці оцінки.

Серед даних, які залишились після етапу відбирання даних з початкової таблиці, є приклади із невідомими значеннями атрибутів. Необхідність опрацювання таких неповних таблиць даних зумовила здійснення двох серій експериментів.

У першій серії експериментів на етапі попереднього опрацювання даних застосовано метод видалення прикладів з невідомими значеннями. Загалом було видалено 18 таких прикладів і розміри таблиці становили 188×38. У другій серії при опрацюванні даних було використано глобальне доповнення, за яким невідомі значення доповнені вибраним параметром. Параметр, яким доповнені відсутні дані, для числових атрибутів обчислений як середнє арифметичне значень відповідного домена, а для символьних – значення їхніх доменів, які зустрічаються найчастіше.

Перша серія експериментів

Таблиця, яку буде використано для аналізу у першій серії експериментів, була опрацьована з використанням методу видалення прикладів з невідомими значеннями атрибутів. Розміри таблиці 188×38. Її структура наведена у табл. 3.

Таблиця 3

Фрaгмент таблиці для першої серії експериментів

Група	*Інформативні*						*Атрибути-характеристики*			*Проміжні оцінки*		*Атрибути рішення*	
Назва атрибута	Професія	Місце роботи	Стать	Освіта	Вікова група	Груп а/стаж	17 атрибутів з оцінками результатів психологічних і психофізіологічних тестів	4 атрибути з оцінками особистих якостей	7 атрибутів з даними про мотиви	Оцінка ІП	Профпридатність	Успішність	Надійність
	5	ВЕМ	ч	1	1	1	[в, д, з, н]	[в, д, з, н]	[0, 1]	д	п	4	5
	5	ВЕМ	ч	1	2	1				з	р	4	4
	5	ВЕМ	ч	1	1	1				д	уп	3	3
	5	ПЕМ	ч	1	1	1				д	бп	4	5
	4	ВЕМ	ч	1	3	1				в	п	4	4
	5	ГЕМ	ч	1	1	2				з	п	4	4
	5	ОЕМ	ж	1	4	4				д	п	4	4
	5	ВЕМ	ч	1	1	1				д	п	4	5
	5	ВЕМ	ч	2	1	1				в	п	5	5
	5	ВЕМ	ч	1	1	3				з	п	4	4
. . .													
	5	ПЕМ	ж	1	4	4				д	п	4	4

Над даними з табл. 3 виконано низку експериментів. Для кожного експерименту було відібрано свій набір атрибутів.

1. Для експерименту Е1.1 відібрано 6 інформативних атрибутів, атрибут „оцінка ІП”, 4 атрибути з оцінками особистих якостей, 7 атрибутів з даними про мотиви, атрибут „профпридатність”, „успішність” та атрибут прийняття рішення – „надійність”. Разом – 21 атрибут. Розміри таблиці прийняття рішень – 188×21.

2. В експеримент Е1.2 увійшли всі 38 атрибутів з початкової табл. 3. На відміну від експерименту Е1.1, в експеримент увійшли 17 атрибутів з результатами психологічних та психофізіологічних тестів, на основі яких розраховується „оцінка ІП”. Атрибут прийняття рішення – „надійність”. Розміри таблиці даних – 188×38.

3. В експерименті Е1.3 відібрано 6 інформативних атрибутів, атрибут „оцінка ІП”, 4 атрибути з оцінками особистих якостей, 7 атрибутів з даними про мотиви, атрибут „профпридатність” та „успішність”. Атрибут прийняття рішення – „успішність”. Загальний розмір таблиці – 188×20.

4. В експеримент Е1.4 увійшли всі атрибути з початкової табл. 3 за винятком атрибута „надійність”. У цьому експерименті таблиця прийняття рішень містить атрибути з результатами

психологічних та психофізіологічних тестів, на основі яких розраховується „оцінка ІП”. Атрибут прийняття рішення – „успішність”. Розміри таблиці даних – 188×37.

5. Для експерименту Е1.5 відібрано 6 інформативних атрибутів, атрибут „оцінка ІП”, 4 атрибути з оцінками особистих якостей, 7 атрибутів з даними про мотиви, атрибут „профпридатність” та „надійність”. Атрибут „успішність” не відібрано з метою перевірки того, чи пов’язана ця суб’єктивна оцінка з іншими атрибутами таблиці. Атрибут прийняття рішення – „надійність”. Загальний розмір таблиці – 188×20.

Побудова редуктів. У табл. 4–8 наведено результати п’яти експериментів (Е1.1-Е1.5), метою яких було знаходження редукта кожної таблиці – набору атрибутів, які найбільше впливають на прийняття рішення. Дослідження здійснювалось засобами системи ROSETTA [9,14], в якій реалізовано методологію наближених множин. Ця система є безкоштовним академічним проектом і містить реалізацію основних алгоритмів опрацювання даних, дискретизації, пошуку редукта, алгоритму класифікації з використанням правил тощо.

Символом "+" у таблицях позначено атрибути, які утворюють редукти за різних параметрів алгоритму.

Таблиця 4

Результати експерименту Е1.1 за методом Джонсона для атрибута прийняття рішення „надійність” (розмір таблиці – 188´21)

Ступінь підтримки, %	Професія	Місце роботи	Вікова група	Оцінка ІП	Потреба досягнень	Активність	Стиль керівництва	Організаційні здібності	Успішність	Мотив – гроші	Мотив – сусп. активність
100	+	+	+	+	+	+	+	+		+	+
98	+	+	+				+	+	+		
95		+	+				+	+	+		
90		+	+				+		+		
81		+	+				+				

Аналіз таблиці прийняття рішень за методом Джонсона дозволив з 21 атрибута, що входили до неї, відібрати до редукта зі ступенем підтримки у 100% десять атрибутів, які наведено у табл. 4. Зі зменшенням ступеня підтримки виявилось, що серед атрибутів, які впливають на прийняте рішення, з’являється „успішність” (у редуктах зі ступенем підтримки 98, 95 та 90%). Найбільше на „надійність” впливають „місце роботи”, „вікова група” та „стиль керівництва”, оскільки ці атрибути входять до редукта зі ступенем підтримки 81%.

Таблиця 5

Результати експерименту Е1.2 за методом Джонсона для атрибута прийняття рішення „надійність” (розмір таблиці – 188´38)

Ступінь підтримки, %	Місце роботи	Оцінка об’єму уваги	Оцінка точності відтворення	Оцінка швидкості відтворення	Оцінка технічного мислення	Стиль керівництва
100	+	+	+	+	+	+
97	+		+	+		
91	+		+			

В експерименті Е1.2 з іншими незмінними умовами було додано 17 атрибутів, на основі яких обчислюється інтеґральний показник. Мета досліду – виявити, чи впливають складові інтеґрального показника на прийняте рішення.

Як і у попередньому експерименті, атрибути „місце роботи” та „стиль керівництва” знову належать до редукта зі 100% підтримкою. Також виявилось, що чотири атрибути, які враховуються при обчисленні інтеґрального показника, є впливовішими, ніж особистісні характеристики та мотиви, які проявилися в експерименті Е1.1. Тому можна зробити висновок, що на „надійність” більше впливають психологічні та психофізіологічні характеристики, ніж „професія”, „вікова група”, „потреба досягнень”, „активність” та „організаційні здібності”.

Таблиця 6

Результати експерименту Е1.3 за методом Джонсона для атрибута прийняття рішення „успішність” (розмір таблиці – 188´20)

Ступінь підтримки, %	Професія	Місце роботи	Вікова група	Оцінка ІП	Потреба досягнень	Активність	Стиль керівництва	Мотив – суєп. активність	Мотив – суєп. корисність
100	+	+	+	+	+	+	+	+	+
98	+	+	+		+	+	+	+	
95		+	+		+	+	+	+	
91		+	+		+	+	+		
84		+			+	+	+		

Таким дослідом мали на меті виявити, які атрибути впливають на „успішність”. До таких алгоритмом було відібрано 10 атрибутів, найсильніші з яких: „місце роботи”, „потреба досягнень”, „активність”, „стиль керівництва”. Також впливовим є атрибут „вікова група”, який входить в редукти зі ступенем підтримки від 91 до 100%.

Таблиця 7

Результати експерименту Е1.4 за методом Джонсона для атрибута прийняття рішення „успішність” (розмір таблиці – 188´37)

Ступінь підтримки, %	Місце роботи	Оцінка об’єму уваги	Оцінка точності відтворення	Оцінка швидкості відтворення	Оцінка логічного мислення	Оцінка стійкості реакції
100	+	+	+	+	+	+
97	+	+		+		
89	+	+				

Цей експеримент аналогічний до Е1.2, але за рішення взято атрибут „успішність”. До редукта увійшли 5 атрибутів, які є складовими інтегрального показника та атрибут „місце роботи”. Бачимо, що як і в результатах досліду Е1.3, „місце роботи” є впливовим атрибутом. Також видно, що

підтверджуються й інші результати аналізу з табл. 6, а саме: вплив компонентів ІП на прийняте рішення є значно сильнішим, ніж вплив атрибутів „професія”, „вікова група”, особистісних характеристик та мотивів.

Таблиця 8

Результати експерименту Е1.5 за методом Джонсона для атрибута прийняття рішення „надійність” (розмір таблиці – 188´20)

Ступінь підтримки, %	Професія	Місце роботи	Вікова група	Група/стаж	Оцінка ІП	Потреба досягнень	Активність	Стиль керівництва	Організаційні здібності	Мотив – гроші	Мотив – сусп. активність	Мотив – комфорт
100		+	+	+	+	+	+		+	+	+	+
97	+	+	+			+	+	+			+	
94		+	+			+	+	+			+	
81		+	+				+	+				

У результаті цього аналізу виявлено, що при вилученні з розгляду атрибута „успішність” на „надійність” впливають ті самі атрибути, що були виявлені в першому досліді (табл. 4) та додатково „група/стаж” і „мотив – комфорт”. Крім цього, посилився вплив атрибута „активність” (порівняно з результатами досліду Е1.1).

Ґенерування правил. На основі знайдених редуктів побудовано різні набори правил. Наведемо лише деякі з них.

На основі редукта {„професія”, „місце роботи”, „вікова група”, „стиль керівництва”, „організаційні здібності”, „успішність”} зі ступенем підтримки 98%, який отримано у експерименті Е1.1, згенеровано 156 правил.

Нижче наведено правила, рішенням у яких є „надійність”=5 (максимальна надійність):

1. професія(5) ∧ місце роботи(ОЕМ) ∧ вікова група(2) ∧ стиль керівництва(в) ∧ організаційні здібності(д) ∧ успішність(4) => надійність(5)
2. професія(5) ∧ місце роботи(ОЕМ) ∧ вікова група(2) ∧ стиль керівництва(д) ∧ організаційні здібності(з) ∧ успішність(4) => надійність(5)
3. професія(4) ∧ місце роботи(ГЕМ) ∧ вікова група(1) ∧ стиль керівництва(в) ∧ організаційні здібності(д) ∧ успішність(4) => надійність(5)
4. професія(5) ∧ місце роботи(ГЕМ) ∧ вікова група(3) ∧ стиль керівництва(д) ∧ організаційні здібності(в) ∧ успішність(5) => надійність(5)
5. професія(4) ∧ місце роботи(ОЕМ) ∧ вікова група(4) ∧ стиль керівництва(в) ∧ організаційні здібності(в) ∧ успішність(4) => надійність(5)

Далі наведено правила, рішенням для яких є „надійність”=3 (мінімальна надійність):

1. професія(5) ∧ місце роботи(ВЕМ) ∧ вікова група(1) ∧ стиль керівництва(в) ∧ організаційні здібності(д) ∧ успішність(3) => надійність(3)
2. професія(5) ∧ місце роботи(ВЕМ) ∧ вікова група(1) ∧ стиль керівництва(з) ∧ організаційні здібності(в) ∧ успішність(4) => надійність(3)
3. професія(2) ∧ місце роботи(ВЕМ) ∧ вікова група(3) ∧ стиль керівництва(д) ∧ організаційні здібності(д) ∧ успішність(4) => надійність(3)

4. професія(2) ∧ місце роботи(ПЕМ) ∧ вікова група(1) ∧ стиль керівництва(з) ∧ організаційні здібності(д) ∧ успішність(4) => надійність(3)

5. професія(5) ∧ місце роботи(ПЕМ) ∧ вікова група(1) ∧ стиль керівництва(д) ∧ організаційні здібності(в) ∧ успішність(4) => надійність(3)

Друга серія експериментів

Для другої серії експериментів було здійснено глобальне доповнення відсутніх значень вибраним параметром. Як параметр для числових значень атрибута обрано середнє, для символьних – значення, що зустрічається найчастіше. У системі ROSETTA цей метод має назву MMF (*mean/mode fill)*.

Перед доповненням з табл. 3 було вилучено 12 прикладів, для яких значення атрибута прийняття рішення було відсутнім. Таблиця прийняття рішень містила 37 умовних атрибутів, тобто 37 атрибутів × 194 прикладів = 7178 значень атрибутів. Серед них було шість прикладів з відсутніми значеннями атрибутів, що становить 3.1% від загальної кількості прикладів; відсутніх значень атрибутів – 57, що становить 0.79%. Як вказано раніше, такий рівень відсутніх значень цілком прийнятний для аналізу.

Потім було заповнено відсутні значення для 6 прикладів. Для 5 з них були заповнені відсутні оцінки особистісних характеристик та мотивів. Для одного прикладу заповнено значення двох атрибутів з оцінками результатів психологічних і психофізіологічних тестів. Остаточні розміри готової для аналізу таблиці прийняття рішень склали 194×38.

На основі цієї таблиці відібрано дані для серії з 5 експериментів та побудовано таблиці з їх результатами. Ці таблиці мають номери 9–13. Вони відповідають експериментам, які позначені як E2.1 – E2.5. Сформовані таблиці даних складаються з тих самих атрибутів, що й відповідні їм таблиці у експериментах E1.1 – E1.5, тобто мають однакову кількість стовпців. Відмінність полягає в тому, що всі нові таблиці даних мають 194 рядки, на відміну від таблиць з першої серії експериментів, які містили 188 рядків.

Побудова редуктів. Далі у табл. 9–13 наведено результати експериментів з побудови редуктів для таблиць з доповненими даними.

Таблиця 9

Результати експерименту E2.1 за методом Джонсона для атрибута прийняття рішення „надійність” (розмір таблиці – 194 ´ 21)

Ступінь підтримки, %	Професія	Місце роботи	Освіта	Вікова група	Група/стаж	Оцінка ІП	Активність	Стиль керівництва	Організаційні здібності	Успішність
100	+	+	+	+	+	+	+	+	+	+
98	+	+		+				+	+	+
95		+		+				+	+	+
91		+		+				+		+
82		+		+				+		

Результати такого аналізу незначно відрізняються від аналогічних результатів першої серії: видно, що найвпливовішими залишилися „місце роботи”, „вікова група” та „стиль керівництва”. Редукт таблиці прийняття рішень з експерименту E2.1 демонструє, що порівняно з результатами

експерименту Е1.1, видаленням прикладів з неповним описом, видалено ті з них, які визначали важливість атрибутів „надійність роботи”, „освіта” та „група/стаж”. Заповнення відсутніх даних підтвердило важливість особистісних характеристик, але показало, що мотиви не впливають на атрибут прийняття рішення.

Таблиця 10

Результати експерименту Е2.2 за методом Джонсона для атрибута прийняття рішення „надійність” (розмір таблиці даних – 194´38)

Ступінь підтримки, %	Місце роботи	Оцінка об’єму уваги	Оцінка точності відтворення	Оцінка швидкості відтворення	Оцінка технічного мислення	Стиль керівництва
100	+	+	+	+	+	+
97	+		+	+		
91	+		+			

У результаті цього досліду було отримано ті самі залежності, що й при аналізі в експерименті Е1.2 (табл. 5). Нові приклади з доповненими значеннями внесли в таблицю додаткову інформацію, яка не спростовує вплив компонентів інтеґрального показника на атрибут прийняття рішення.

Таблиця 11

Результати експерименту Е2.3 за методом Джонсона для атрибута прийняття рішення „успішність” (розмір таблиці даних – 194´20)

Ступінь підтримки, %	Професія	Місце роботи	Стать	Освіта	Вікова група	Оцінка ІП	Потреба досягнень	Активність	Стиль керівництва	Мотив – гроші	Мотив – сусп. активність
100	+	+	+	+		+	+	+	+	+	+
98	+	+			+		+	+	+		+
94	+	+			+		+	+	+		
90		+			+		+	+	+		
82		+			+			+	+		

Порівнюючи результати цього експерименту з результатами у табл. 6 бачимо, що проявився вплив атрибутів „стать” і „освіта”. Це корелює з результатами аналізу з табл. 9, де з’являється вплив „освіти” та „групи/стажу”. Очевидно, приклади, що були видалені у першій серії експериментів, містять дані, які підсилюють значимість атрибута „освіта”.

Порівняно з результатами першої серії експериментів (табл. 7) результати експерименту Е2.4 не показують значних відмінностей. Після доповнення даних замість атрибута „оцінка логічного мислення” важливішим виявився атрибут „потреба досягнень”.

Таблиця 12

Результати експерименту Е2.4 за методом Джонсона для атрибута прийняття рішення „успішність” (розмір таблиці даних – 194´37)

Ступінь підтримки, %	Місце роботи	Оцінка об’єму уваги	Оцінка точності відтворення	Оцінка швидкості відтворення	Оцінка стійкості реакції	Потреба досягнень
100	+	+	+	+	+	+
97	+	+		+		
90	+	+				

Таблиця 13

Результати експерименту Е2.5 за методом Джонсона для атрибуту прийняття рішення „надійність” (розмір таблиці даних – 194´20)

Ступінь підтримки, %	Професія	Місце роботи	Вікова група	Оцінка іп	Потреба досягнень	Активність	Стиль керівництва	Організаційні здібності	Мотив – гроші	Мотив – суєп. активність
100	+	+	+	+	+	+	+	+	+	+
97	+	+	+		+	+	+			+
94	+	+	+			+	+			+
82		+	+			+	+			

Порівнюючи ці результати з результатами у табл. 8, бачимо, що після доповнення підсилилась важливість „професії”, а атрибути „група/стаж”, „мотив – комфорт” зникли з редукта. Порівнюючи результати у табл. 13 і табл. 9, можна зазначити, що із видаленням з таблиці прийняття рішень атрибута „успішність” з редукта зникають також і атрибути „освіта” і „група/стаж”. Натомість важливість атрибута „активність” підсилюється. Отже, можна зробити висновок, що „успішність” „освіта” та „група/стаж” взаємопов’язані. Це підтверджує аналіз результатів експерименту Е2.3, які показують, що „освіта” впливає на „успішність” у випадку редукта зі 100% підтримкою.

Також, з результатів, наведених у табл. 3, 6, 8, 9, 11, 13, можна прослідкувати взаємозв’язок атрибутів „активність” та „успішність”. Зокрема, за наявності у таблиці прийняття рішень атрибута „успішність”, вплив атрибута „активність” незначний (він входить лише в редукт зі 100% підтримкою). При видалені з таблиці атрибута „успішність”, її роль у залежності посідає атрибут „активність”, який стає одним із найвпливовіших (табл. 8 і табл. 13). Крім цього, „активність” завжди була сильним атрибутом у випадках, коли атрибутом прийняття рішення обирали „успішність” (табл. 6 і табл. 11).

Аналіз результатів, отриманих у двох серіях експериментів

Одним з головних результатів дослідження стало виявлення набору атрибутів, які найбільше впливають на прийняте рішення щодо успішності та надійності роботи працівника. Поділимо ці атрибути на більш важливі та менш важливі. Важливішими вважатимемо ті атрибути, які входять до “сильніших” редуктів зі ступенем підтримки 98% і менше. Менш важливими вважатимемо атрибути, які входять лише до редуктів з підтримкою 100%.

До найважливіших ознак віднесемо такі 12 атрибутів: „професія”, „місце роботи”, „вікова група”, „оцінка об’єму уваги”, „оцінка точності відтворення”, „оцінка швидкості відтворення”, „потреба досягнень”, „активність”, „стиль керівництва”, „організаційні здібності”, „мотив – суспільна активність”, „успішність”.

До менш важливих ознак потрапили такі 10 атрибутів: „стать”, „освіта”, „група/стаж”, „оцінка технічного мислення”, „оцінка логічного мислення”, „оцінка стійкості реакції”, „оцінка ІП”, „мотив – гроші”, „мотив – суспільна корисність”, „мотив – комфорт”.

Те, що атрибут „оцінка ІП”, який фактично є проміжним рішенням і значення якого розраховується на базі 17 атрибутів, виявився менш важливою ознакою, пояснюємо недосконалою методикою його розрахунку. Очевидно, ця оцінка недостатньо відображає залежності між атрибутами, на основі яких вона розраховується, та не впливає на залежності у всій таблиці.

Наступні 15 атрибутів не увійшли до жодного редукта: „оцінка концентрації”, „оцінка переключення уваги”, „оцінка кількості помилок”, „оцінка швидкості технічного мислення”, „оцінка оперативної пам’яті”, „оцінка об’єму оперативної пам’яті”, „оцінка короткочасної пам’яті”, „оцінка швидкості реакції”, „оцінка рівня функціональних можливостей”, „оцінка сили нервової системи”, „оцінка рухливості нервових процесів”, „мотив – соціальний статус”, „мотив – творча активність”, „мотив – спілкування”, „профпридатність”. Отже, характеристики, що відповідають вказаним атрибутам, незначно впливають на успішність та надійність роботи працівників або взагалі не впливають.

Під час досліджень виявлено, що доповнення відсутніх значень методом MMF певною мірою впливає на кінцеві результати. З порівняння результатів аналізу даних з доповненням і без доповнення видно, що навіть незначний обсяг (0.79%) доповнених значень дає змогу дещо покращити якість отриманих результатів. Зокрема, у першій серії експериментів, у табл. 3 і табл. 8 спостерігається деяка суперечливість: у редуктах зі 100% підтримкою відсутні декілька атрибутів („успішність” у табл. 3; „професія”, „стиль керівництва” у табл. 8), які далі з’являються у редуктах з меншим ступенем підтримки.

Це може означати, що в таблиці даних є декілька прикладів, які призводять до того, що алгоритм Джонсона спочатку знаходить залежність, яка виконується для всієї таблиці, але не є „правильною”. Тільки зі зменшенням ступеня підтримки редукта знаходиться „правильна” залежність, яка, очевидно, не поширюється на ці „проблемні” приклади.

У випадку ж доповнення відсутніх значень кількість прикладів у таблиці на 6 більша і становить 194. Результати побудови редуктів показують, що суперечність виникає лише в одному експерименті – у табл. 11 видно, що до редукта зі 100% підтримкою не увійшов атрибут „вікова група”, який, в принципі, є важливою ознакою, оскільки входить до редуктів з підтримкою 98% і нижче.

З іншого боку, набори атрибутів, знайдені для заповненої таблиці, дещо відрізняються від знайдених для незаповненої. Це можна пояснити порівняно невеликою кількістю – 188 – повністю описаних прикладів у таблиці прийняття рішень у першій серії експериментів, через що поява шістьох додаткових прикладів у другій серії експериментів вносить нові залежності між атрибутами.

Висновки та подальший напрямок досліджень

Основною метою і основним результатом дослідження стало виявлення набору атрибутів, які найбільше впливають на прийняте рішення щодо успішності та надійності роботи працівника. На основі цих результатів надалі можна буде більш обґрунтовано відбирати окремі, лише вагомі, атрибути для експериментів над даними чи для здійснення майбутніх тестувань персоналу. Виявлено, що із 20–38 (в різних випадках) атрибутів тільки 6–10 атрибутів впливають на прийняте рішення.

У процесі опрацювання таблиць даних потрібно було врахувати відсутність частини даних. Тому було здійснено дві серії експериментів, у кожній з яких застосовано окремий спосіб врахування відсутніх даних.

Заповнено відсутні значення і генерувано правила, проте повне оцінювання результатів цих дій вимагає подальших досліджень. Надалі важливо проаналізувати, який спосіб опрацювання невідомих значень для такого типу даних дає кращі результати. Необхідно порівняти результати, отримані після доповнення даних різними методами та обрати оптимальний.

Використовуючи різні методи доповнення, необхідно докладніше проаналізувати залежність, виявлену між атрибутами „освіта" та „успішність" в другій серії експериментів після застосування методу глобального доповнення. Також потрібно докладніше дослідити взаємний вплив атрибутів „активність" та „успішність", який виявився в процесі аналізу даних.

У таблиці прийняття рішень існує атрибут віку працівника та атрибут стажу. Для цих атрибутів треба провести дискретизацію значень. Це дозволить утворити нові атрибути з даними про групи працівників за віком та за стажем, які повинні краще відображати залежності у таблиці, ніж наявні атрибути „вікова група" та „група/стаж". Також потрібно оцінити чинники, які впливають на професійне довголіття – йдеться про аналіз профпридатності груп, утворених дискретизацією.

Варто проаналізувати, які з атрибутів найбільше впливають на прийняте рішення, якщо дослідження проводити для окремих груп – за освітою (середня спеціальна, вища), статтю, місцем роботи, посадою (три ланки керівників різного рівня), віком та стажем. Потрібно дослідити, які з характеристик, у першу чергу – особистісні характеристики та оцінки знань працівника, не введені в інтегральний показник, впливають на нього.

1. Mitra Sushmita Data mining in soft computing framework: a survey / Mitra Sushmita, Pal Sankar K., Mitra Pabitra // IEEE Transactions on Neural Networks. – 2002. – Vol. 13. – Issue 1. 2. Модели и алгоритмы принятия решений в нечетких условиях [Електронний ресурс] / А. Е. Алтунин, М. В. Семухина // Тюмень : Изд-во Тюменского государственного университета, 2000. – Режим доступу до журн. : http://sedok/narod/ru/s_files/tumen.htm. 3. Acuna E. The Treatment of Missing Values and its Effect in the Classifier Accuracy / E. Acuna, C. Rodriguez // Classification, Clustering and Data Mining Applications, Springer-Verlag. – Berlin-Heidelberg, 2004. – C. 639-648. 4. Jan Komorowski Rough Sets: A Tutorial / Jan Komorowski, Lech Polkowski, Andrzej Skowron // Eds. S.K. Pal and A. Skowron, Rough Fuzzy Hybridization: A New Trend in Decision-Making, Spriner-Verlag. – Singapore, 1998. 5. Grzymala-Busse J.W. Rough Set Strategies to Data with Missing Attribute Values / J. W. Grzymala-Busse // Foundation and New Directions in Data Mining : Workshop; Data Mining : Proceedings on third IEEE International Conference. – Melbourne, FL, USA, 2003. – C. 56-63. 6. Metody wnioskowania w oparciu o niekompletny opis obiektow [Електронний ресурс] / R. Latkowski // Warszawa, 2001. – Режим доступу до журн. : http://logic.mimuw.edu.pl/Grant2003/prace/BMscLatkowski.pdf. 7. RSES 2.2. User's Guide [Електронний ресурс] / Режим доступу до журн. : http://logic.mimuw.edu.pl/~rses/RSES_doc_end/pdf. 8. The Rule Induction System LERS – a Version for Personal Computers [Електронний ресурс] / M. R. Chmielewski, J. W. Grzymal-Busse, N. W. Peterson // Режим доступу до журн. : http://coitweb.uncc.edu/~ras/KDD-02/LERSforPC.pdf. 9. ROSETTA Technical Reference Manual [Електронний ресурс] / A. Øhrn, 2001 // Режим доступу до журн. : http://www.idi.ntnu.no/~aleks/. 10. Pawlak Z. Rough Sets – Theoretical Aspects of Reasoning about Data / Z. Pawlak // Series D: System Theory, Knowledge Engineering and Problem Solving. – volume 9 of. Kluwer Academic Publishers, Dordrecht, 1991. 11. Data Mining and Rough Set Theory [Електронний ресурс] / J. W. Grzymala-Busse, W. Ziarko // Communications of the ACM – 2000. – Vol. 43. – №4. – Режим доступу до журн. : http://www.csc.ncsu.edu/faculty/mpsingh/papers/columns/cacm-00-anykey.pdf. 12. Grzymala-Busse J.W. MLEM2 – Discretization During Rule Induction / J. W. Grzymala-Busse // IIPWM'2003, Intelligent Information Processing and WEB Mining Systems : Proceedings on the International Conference, June 2-5 2003. – Zakopane, Poland, 2003. – C. 499-508. 13. Konias S. A Novel Approach for Incremental Uncertainty Rule Generation from Databases with Missing Value Handling: Application to Dynamic Medical Databases / S. Konias, I. Chouvarda, I. Vlahavas, N. Maglaveras // Medical Informatics & The

Internet in Medicine. – *Taylor & Francis, 2005. – Vol. 2005, Issue 5. 14. Øhrn A. Discernibility and Rough Sets in Medicine: Tools and Applications : PhD thesis, Norwegian University of Science and Technology, Department of Computer and Information Science / A. Øhrn. – 1999. 15. Нікольський Ю.В., Щербина Ю.М. Генетичні алгоритми в екстремальних задачах / Ю.В. Нікольський, Ю.М. Щербина // Вісник Львівського університету, Серія прикладна математика та інформатика. – 2000. – Вип. 2. – С. 191–208.*

УДК 004.93.1

В. М. Заяць[*], М.М. Заяць
[*]Львівський державний інститут новітніх
технологій та управління імені В. Чорновола,
кафедра інформаційно-комп'ютерних технологій та систем;
Національний університет „Львівська політехніка",
кафедра інформаційних систем і мереж

ПІДХОДИ ДО ПОБУДОВИ СИСТЕМ ЗАХИСТУ ІНФОРМАЦІЇ ВІД НЕСАНКЦІОНОВАНОГО ДОСТУПУ

Розглянуто методи побудови систем захисту інформації від стороннього доступу. Запропоновано підхід до побудови системи захисту комп'ютера на основі використання автоматизованої системи розпізнавання та ідентифікації користувача комп'ютера, що ґрунтується на вимірюванні часових затримок при введенні інформації з клавіатури комп'ютера у дискретні відліки часу. Це дало змогу підвищити ефективність розпізнавання майже в 1,5 раза, автоматизувати процедуру ідентифікації користувачів, забезпечити надійний захист інформації та гарантувати конфіденційний доступ до неї.

In the article the proposed methods to building of the defense system of information from strange access on base of system of computer users recognition with help of model discrete. It is made by measure of time delays at the entered information from a keyboard of computer in the discrete moments of time. It allowed to increase efficiency of recognition almost in 1,5 times and to automate procedure of users authentication and defense of information from strange access created on the basis of discrete models are marked.

Постановка проблеми

При створенні реальних систем захисту інформації від стороннього доступу, дослідженні фізичних явищ чи процесів, побудові систем розпізнавання, ідентифікації та зберігання інформації з бажаними характеристиками доцільно провести їх аналіз та комп'ютерне моделювання шляхом створення математичної моделі системи, що розробляється. Такий підхід вимагає значно менших часових і технічних засобів порівняно з фізичним експериментом, особливо на попередній стадії розробки, коли системи чи пристрою, що розробляється, немає.

Останнім часом в нелінійній динаміці широке застосування знаходять дискретні моделі систем [1–5], для яких дискретність закладена в природі самого об'єкта досліджень, а не є наслідком дискретизації неперервної системи. Доцільність використання дискретних за своєю природою моделей пояснюється такими їх особливостями: