УДК 519.226

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АНАЛІЗУ ФІНАНСОВИХ ДАНИХ НА ОСНОВІ ІНТЕГРОВАНОГО МЕТОДУ

Н.В. КУЗНЄЦОВА, П.І. БІДЮК

Проаналізовано основні особливості фінансових даних та запропоновано новий інтегрований метод їх аналізу. Запропоновано нову інформаційну технологію на основі інтегрованого методу аналізу даних та на практиці проілюстровано можливість її застосування для оцінювання кредитоспроможності позичальника.

ВСТУП

Мінливість та нестабільність розвитку сучасного світу, великі обсяги інформації в різних галузях науки, які необхідно обробляти з метою прийняття коректних рішень, спонукають до пошуку нових методів та підходів для опрацювання даних. Системну невизначеність, яка дедалі частіше наявна в даних, необхідно обробляти, знаходити певні закономірності та фактори впливу. Для виявлення взаємозв'язків між окремими змінними задачі використовують математичні методи регресійного аналізу (логістичної регресії), дерева рішень, мережі Байєса, нейронні мережі, кластерний аналіз, нечітку логіку тощо [1, 2, 3, 4]. Однак, незважаючи на наявність достатньої множини методів аналізу даних, не завжди вдається отримати бажаний (прийнятний) результат щодо розв'язання конкретних задач обробки даних та прийняття рішень. Тому необхідно удосконалювати існуючі методи, розробляти нові, а також комбінувати різні підходи для досягнення бажаної точності результату.

ПОСТАНОВКА ЗАДАЧІ

Мета роботи — проаналізувати особливості фінансових даних та існуючих методів для їх аналізу, запропонувати новий метод аналізу даних, який передбачає комбінацію існуючих підходів та на практичному прикладі проілюструвати ефективність застосування інтегрованого методу аналізу даних.

У роботі пропонується логічна організація процесу збору й аналізу фінансових даних, а також запропоновано нову інформаційну технологію на основі інтегрованого методу та розглянуто особливості її застосування на практиці.

ПРОБЛЕМИ АНАЛІЗУ ДАНИХ У ФІНАНСОВИХ УСТАНОВАХ

На сьогодні фінансові установи використовують різноманітні програмні продукти для аналізу даних. Це найбільш відомі зарубіжні системи SAS, SAP, SPSS і власні розробки програмістів й аналітиків фінансових установ. Найзручнішим для використання вважається той програмний продукт, який

не потребує додаткових інструментів обробки чи аналізу даних, знань та навичок від користувачів. Тому вони досить скептично ставляться до запровадження в експлуатацію нових інструментів доти, поки звичні інструментальні засоби продовжують працювати. Більшість із таких програмних продуктів ґрунтуються на одному або декількох відомих методах аналізу даних; при цьому найчастіше використовуються такі: логістична регресія, нейронні мережі, дерева рішень та мережі Байєса (МБ) — новітній інструмент ймовірнісного аналізу даних.

Логістична регресія — це вид нелінійної множинної регресії, яка аналізує функціональну залежність між декількома незалежними змінними (регресорами) і залежною змінною [2, 5]. Бінарна логістична регресія застосовується у тому випадку, коли вихідна змінна може приймати тільки два значення.

У множинній регресії припускається, що залежна змінна є лінійною функцією незалежних змінних, тобто: $y = b_1 x_1 + b_2 x_2 + ... + b_n x_n + u$, де y — залежна змінна (результат прийняття рішення); x_i — пояснююча змінна (критерій); b_i — вага пояснюючої змінної i; u — випадкова помилка, $P(u) = 0$. У векторному вигляді це може бути записано таким чином: $y = b'x + u$, де x — вектор пояснюючих змінних, а b' — транспонований вектор параметрів пояснюючих змінних [1]. Відповідно, умовна ймовірність події обчислюється за виразом: $P(y \mid x) = b'x$.

Недоліком логістичної регресії є те, що вона застосовується лише до обмеженої кількості вхідних факторів, тобто на етапі підготовки даних необхідно або додатково залучати експертів, або проводити додаткову обробку даних для виявлення найвпливовіших характеристик і включення в модель лише їх. Можуть також виникнути проблеми з аномальними даними, а також інколи з'являється необхідність відкидання викидів, регуляризації ваг, відкидання ознак, стандартизації даних. Трудомісткість методу вища за звичайний метод найменших квадратів, оцінки ймовірностей можуть виявитися неадекватними, якщо функція правдоподібності не експоненціального вигляду тощо.

Інший метод аналізу фінансових даних, який є досить відомим і поширеним на практиці — дерева рішень. Це один із методів автоматичного аналізу даних, коли правила представляються у вигляді послідовної ієрархічної структури, в якій кожному вузлу відповідає єдиний вузол, що генерує рішення. Під правилом розуміють конструкцію, яка представлена у вигляді «*якщо* ... , *то*...» [2, 3].

Перевагами застосування дерев рішень є такі: швидкий процес навчання, генерування правил у тих галузях, де знання складно формалізувати, зрозуміла класифікаційна модель, висока точність прогнозу. Однак їх застосування пов'язане, також, із низкою суттєвих недоліків. Зокрема, коли виникає необхідність реалізації навчання в оперативному режимі, існуючі алгоритми виявляються занадто громіздкими і потребують значних обсягів пам'яті.

МБ — це ймовірнісна модель у вигляді спрямованого ациклічного графу, кожний вузол якого представляє змінну модельованого процесу, а кожна дуга представляє причинне відношення (зв'язок) між двома змінними [4]. Змінні характеризуються розподілом ймовірності для кожного значення. На

розподіл ймовірності кожного вузла впливають стани (для дискретних вузлів) або значення (для неперервних вузлів) кореневої вершини. Умовні ймовірності станів вузлів зберігаються у таблицях умовних ймовірностей.

Формально МБ — це трійка $N =< V, G, J >$, першою компонентою якої є множина змінних V; другою — спрямований ациклічний граф G, вузли якого відповідають випадковим змінним модельованого процесу, а J — спільний розподіл ймовірностей змінних $V = \{X_1, X_2, ..., X_n\}$. При цьому виконується марковська умова: кожна змінна мережі не залежить від усіх інших змінних, за винятком батьківських попередників цієї змінної.

МБ дозволяє поєднати просте графічне представлення певного процесу з його імовірнісним характером, проаналізувати можливі варіанти розвитку ситуації, відстежити правильність встановлення причинно-наслідкового зв'язку між окремими подіями і завдяки цьому підвищити обґрунтованість рішення під час складних проблемних ситуацій. Основні труднощі, які необхідно подолати при побудові та застосуванні МБ — це побудова самої причинно-наслідкової моделі (первинної структури мережі), суб'єктивність експертів під час визначення апріорних ймовірностей, автоматизація процесів побудови та навчання мережі, забезпечення належної обчислювальної ефективності. За відносно короткий проміжок часу (близько 15 років) МБ вже знайшла успішне застосування при розв'язанні задач медичної та технічної діагностики, кластеризації, розпізнавання та ймовірнісного прогнозування.

ОСОБЛИВОСТІ ФІНАНСОВИХ ДАНИХ ТА ФІНАНСОВИХ ЗАДАЧ

Зазвичай фінансовими даними називають великі обсяги статистичної інформації щодо фінансового стану підприємства, рівня продажів компанії, відношення попиту та пропозиції і т.ін. Фінансові дані характеризуються надзвичайно великою кількістю характеристик (величин), необхідних для їх опису. Дані надходять із різних джерел у різноманітному вигляді, а тому виникає потреба у досить великому репозитарії для їх збереження і достатньо серйозних інструментах для їх обробки. У зв'язку з тим, що дані надходять із різних джерел, у різних вимірах та одиницях, вони є досить розрізненими і не можуть бути оброблені однією людиною — працівником банку. Тому постає питання автоматизації процесу обробки та аналізу даних, раціоналізації їх та приведення результатів до простого і зрозумілого для користувача вигляду. Фінансові дані можна визначити за такими характеристиками, як періодичність, однотипність, множинність і можливість неоднозначного трактування. Дані можуть містити пряме відношення або опис певного фінансового процесу, а також потребують ретельного збору, перевірки та прогнозування.

На сьогодні аналіз даних і прогнозування очікуваних подій на наступні періоди є досить непростим завданням, для розв'язання якого залучаються різноманітні засоби, — статистичні та аналітичні — що ґрунтуються на математичних методах, будуються певні моделі, встановляються взаємозалежності та взаємозв'язки між окремими змінними. Останнім часом обсяг даних, що необхідно проаналізувати, постійно зростає, і тому інколи немож-

ливо ефективно застосувати ці підходи. Виникає потреба у методі, який дозволить виокремити з-поміж усієї множини даних саме ті, які безпосередньо впливають на результуючу прогнозну величину або сукупність величин.

ІНТЕГРОВАНИЙ МЕТОД АНАЛІЗУ ДАНИХ

Основна ідея інтегрованого підходу полягає в тому, що він передбачає комбінацію відомих методів таким чином, щоб уникнути описаних вище недоліків і працювати у тих випадках, коли інші методи не можуть бути застосовані. Очевидно, що під час побудови прогнозної моделі постає питання, як формалізувати зібрані фінансові дані та виявити, які саме з них є суттєвими. Для цього пропонується спочатку побудувати мережу Байєса, яка встановить причинно-наслідкові зв'язки між змінними, що відповідають факторам, визначить силу зв'язків між цими змінними, а також дозволить виявити змінні, які взагалі не пов'язані з результуючою подією («висячі змінні») [5, 6]. На основі побудованої мережі і встановлених зв'язків можна суттєво скоротити кількість факторів, які слід включати на наступному етапі під час побудови моделі. Відомо, що для логістичної регресії зменшення кількості факторів, які включаються в модель, зазвичай спричиняє погіршення якості моделі. Тому необхідно застосовувати мережу Байєса лише як інструмент зменшення кількості факторів, які будуть включені в модель, а не як інструмент, який виявить найсуттєвіші фактори, а всі інші відкине.

Узагальнений алгоритм реалізації інтегрованого методу аналізу даних

Етап 1. Збір статистичних даних, які мають відношення до задачі, що вирішується.

Етап 2. Формалізація зібраних даних і виявлення, які з них є суттєвими. На цьому кроці будується та навчається мережа Байєса на основі статистичних даних, яка і виявляє суттєві змінні та причинно-наслідкові відношення між ними. Завдяки цьому на наступному кроці при побудові моделі можна скоротити кількість факторів, які необхідно враховувати.

Етап 3. Визначена множина суттєвих факторів та змінні, що їм відповідають, включаються у модель, яка будується на основі відомого методу (логістичної регресії, дерева рішень, кластерного аналізу тощо).

Етап 4. Аналіз отриманих результатів, перевірка якості моделі. У разі прийнятної якості моделі використання моделі для прогнозування даних.

Етап 5. На основі побудованої моделі оцінюється прогноз даних та видача рекомендацій щодо поставленої проблеми.

На основі запропонованого алгоритму можна побудувати множину інтегрованих моделей, серед яких необхідно вибрати кращу для цієї проблемної ситуації та поставленої задачі.

Інтегрована модель на основі мережі Байєса та дерева рішень (ІМБД) — це модель, побудована на основі комбінації двох методів — мережі Байєса та дерева рішень, де на першому кроці для скорочення кількості змінних застосовується мережа Байєса, а для оцінки ймовірності дефолту використовується дерево рішень. Іншою моделлю на основі інтегрованого методу є інтегрована модель на основі логістичної регресії і мережі Байєса (ІМЛБ) —

модель, побудована на основі комбінації методів логістичної регресії та мережі Байєса. На першому кроці будується мережа Байєса, яка визначає суттєві змінні, а на основі суттєвих змінних на другому кроці будується логістична регресія [5, 6].

Описаний вище метод можна узагальнити на випадок застосування на другому кроці іншого методу для виявлення суттєвих факторів, наприклад, логістичної регресії або дерева рішень, а на третьому кроці — для побудови моделі використовувати мережі Байєса.

Інтегрована модель на основі дерев рішень та мереж Байєса. Якщо кількість факторів, що впливають на ключову змінну невелика, то можна застосувати запропонований інтегрований метод «backward». Тобто, для задачі будується дерево рішень, яке встановлює, які змінні безпосередньо мають вплив на результат, а потім ця інформація застосовується при побудові мережі Байєса. Під час побудови структури мережі вона може бути задана повністю або частково із використанням експертних даних. Після цього продовжується побудова структури мережі й у результаті отримується остаточна структура мережі, яка відображає причинно-наслідкові зв'язки між змінними. Слід зауважити, що застосовуючи цей підхід не можна «блокувати» зв'язки між змінними, якщо навіть їх не виявлено деревом рішень, оскільки дерево рішень не дозволяє досягти глибокого розуміння причинно-наслідкових зв'язків між змінними.

Визначимо місце і порядок використання інтегрованого методу аналізу даних у загальній структурі аналізу фінансових даних. Процедуру такого аналізу можна представити як низку етапів, що узагальнюють основні операції обробки фінансових даних (рис. 1).

Перший етап зазвичай реалізується великими компаніями, банками, фінансовими установами за допомогою чіткої організаційної структури філіалів, офісів, торгівельних представників, дилерів тощо. Надану інформацію всіма зазначеними установами будемо називати інформацією з «полів». Цей термін є зрозумілим і відображає лише місце, звідки надходить інформація до головного офісу, тобто від «поля» (низу) структури до головного («верхнього») офісу. Ці структурні організації на місцях збирають статистичну інформацію у вигляді затвердженої певним чином звітності про фінансові показники компанії, рівні продажів, фінансовий стан та дані клієнтів, рівні курсів валют і т.ін. Найчастіше у всіх установах є свої розроблені і затверджені головним офісом однотипні форми звітності — як у паперовому, так і в електронному вигляді. У паперовому вигляді — це форми, анкети, бланки, які заповнюються працівниками в «полях», а вже потім ці дані переносяться та передаються у вигляді електронних файлів. Вимога до цього файлу має бути такою, щоб частина інформації була недоступною для редагування користувачами на місцях (заблокувати можливість зміни порядку та назви полів форм), а частина полів має бути відкрита для запису, тобто для введення необхідної інформації з «полів». На цьому етапі обов'язково має бути перевірено та проконтрольовано коректність введення даних на місці, щоб не допустити великої кількості помилок при внесенні даних в електронний документ.

На другому етапі має бути забезпечено збереження конфіденційності інформації, неможливість втручання сторонніх осіб у процес передачі даних

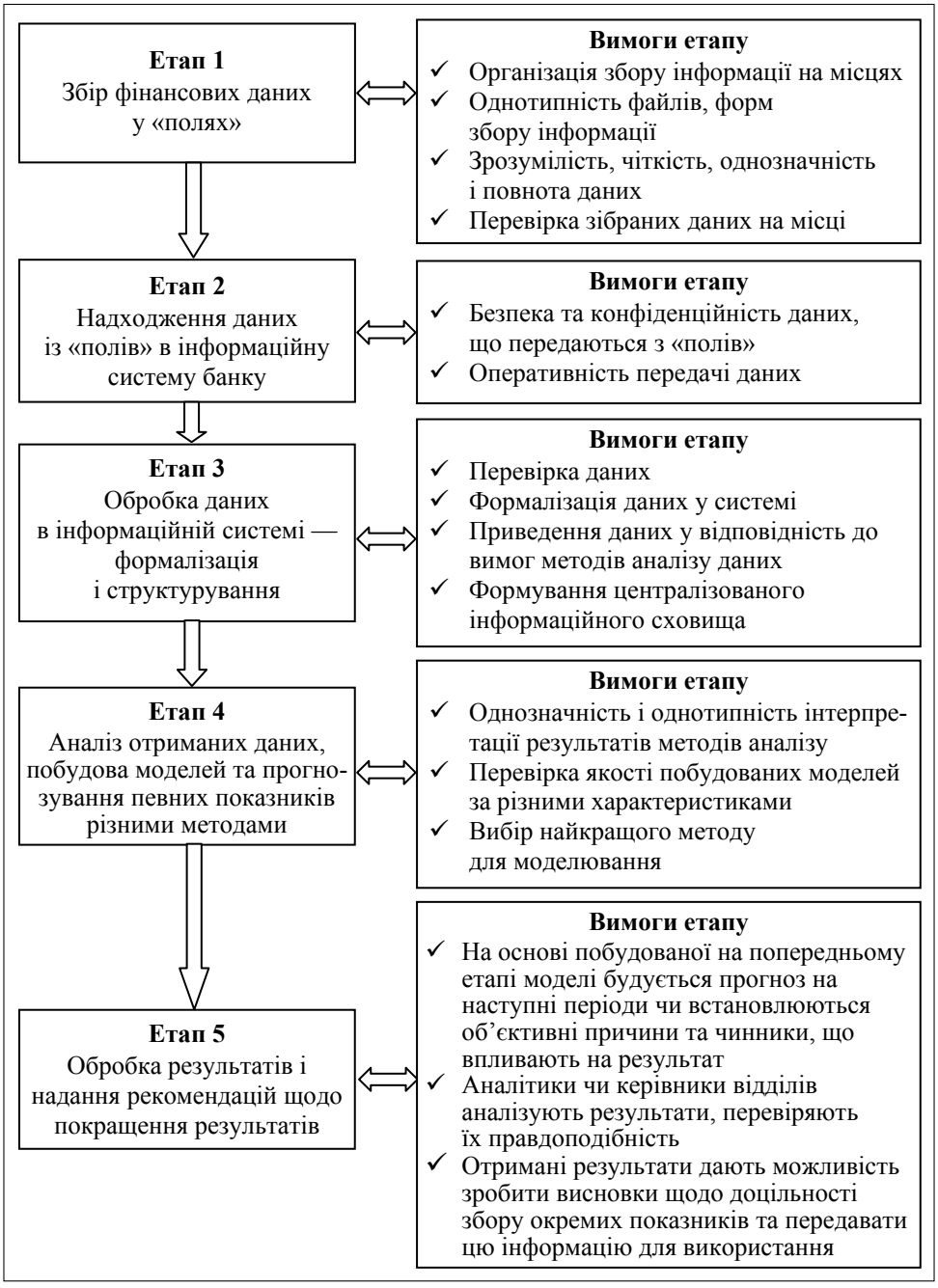


Рис. 1. Організація процесу аналізу фінансових даних

із метою уникнення спотворення чи крадіжки інформації. Ще однією вимогою даного етапу є оперативність передачі інформації до центрального офісу. Оскільки частина фінансових даних збирається під час роботи з клієнтами, то зрозуміло, що час очікування має бути зведений до мінімуму, тобто необхідно, щоб усі процедури введення інформації менеджерами були максимально автоматизовані, а час реакції на їх дії має бути мінімальним. Тому при експлуатації інформаційної технології на практиці мають бути застосо-

вані швидкі та захищені канали передачі інформації від «полів» до центральної інформаційної системи.

На третьому етапі здійснюється обробка даних, що надійшли з підрозділів, перевірка їх коректності, повноти цих даних та приведення до однотипного формалізованого вигляду. Перевірені і формалізовані дані передаються в централізовану систему збереження інформації — інформаційне сховище (базу даних). Слід зазначити, що частина даних у компаніях може зберігатися у таких спеціальних програмних продуктах як SAP, SAS, Nielsen чи на сервері, звідки дані можуть бути легко вивантажені у зручній формі. Дані, що завантажуються в інформаційне сховище, мають бути (але не обов'язково) приведені до зручного вигляду для подальшого аналізу. Якщо дані перед завантаженням не було оброблено, необхідно вивантажити не підготовлені до аналізу дані, привести їх у відповідний вигляд, а потім вже використовувати для подальшого аналізу.

На четвертому етапі наявні підготовлені для аналізу дані. Використовуються доступні методи аналізу даних (логістична регресія, дерева рішень, нейронні мережі, мережі Байєса тощо) та будується модель. Перевіряється адекватність побудованої моделі за різними якісними характеристиками, наприклад, загальна точність, помилки 1-го та 2-го роду. Для цього використовується вибірка, зібрана в інформаційній системі за попередні періоди, яка розбивається на навчальну та перевірочну. Якщо є можливість використати різні методи аналізу даних, то будуються і перевіряються моделі за всіма можливими методами. Серед побудованих моделей обирається краща (або 2–3 кращих) на основі згаданих вище критеріїв.

На п'ятому етапі відібрані на попередньому етапі кращі моделі аналізу даних використовуються для обчислення оцінок прогнозів на наступні періоди. Для уточнення оцінок прогнозів рекомендується комбінувати прогнози, отримані за декількома моделями. При використанні мереж Байєса є можливість виявити причинно-наслідкові зв'язки та встановити причини і чинники, що найкраще впливають на остаточний результат. Виконується обробка результатів, отриманих за певний період на основі побудованої моделі. Аналітики чи керівники відділів надають рекомендації щодо правдоподібності й адекватності отриманих результатів, доцільності збору та використання певних показників на етапі побудови та використання цієї моделі. У разі виявлення недоцільності збору окремих показників чи необхідності зміни їхнього формату, наступна інформація передається в «поля» у вигляді нових файлів та форм збору даних для швидкого застосування їх на практиці і внесення вже оновленої інформації у сховище даних. У межах описаної вище організації аналізу даних компанії, фінансові установи, банки розроблюють чи застосовують відомі технології аналізу даних.

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АНАЛІЗУ ДАНИХ ПОЗИЧАЛЬНИКА БАНКУ

Запропонована інформаційна технологія представляє собою сукупність методів, програмних і технічних засобів, об'єднаних в єдиний технологічний ланцюг, що забезпечує збір, збереження, редагування, обробку, виведення та розповсюдження інформації.

Інформаційна технологія аналізу даних позичальника банку містить модулі збору та збереження інформації (база даних клієнтів); модуль обробки та перевірки інформації (з можливістю залучення експертів-спеціалістів банку); модуль оцінки даних (розробка моделі для аналізу даних) та модуль виведення інформації у вигляді оцінки ймовірності кредиту, або повідомлення про можливість надання кредиту позичальнику (рис. 2).

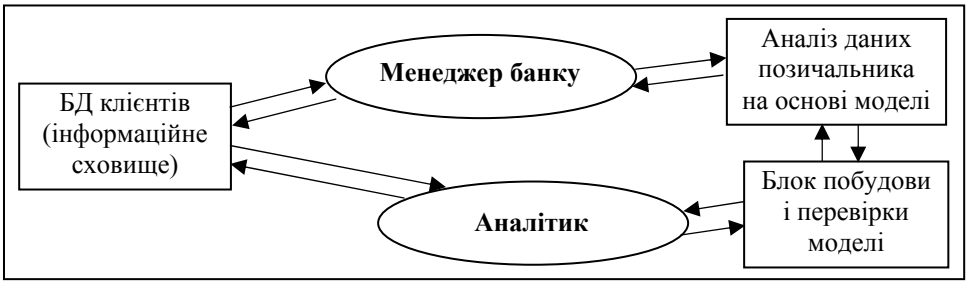


Рис. 2. Основні елементи інформаційної технології аналізу даних позичальника

Така технологія передбачає оцінку даних позичальника, перевірку його кредитоспроможності на основі запропонованого вище інтегрованого методу. Передбачається, що відділення банку видають кредити, збирають усю множину даних — фінансовий стан, соціально-демографічні характеристики позичальника на основі розроблених і встановлених у банку форм-анкет (кредитних заявок). Надані позичальником дані ретельно перевіряються менеджером кредитного відділу під час обробки кредитних заявок перед прийняттям рішення щодо видачі кредиту, уточнюються та вводяться в інформаційну систему банку. Крім цього, в систему вводиться інформація щодо суми кредиту, процентної ставки, дати видачі, строку кредиту. В процесі життєвого циклу кредиту (тобто протягом терміну обслуговування кредиту в банку до моменту його сплати) вноситься інформація щодо вчасності та повноти внесень щомісячної оплати кредиту, а наприкінці строку обслуговування кредиту відмічається (нулем або одиницею), тобто чи відбувся дефолт. Таким чином, банком збирається база позичальників банку, з якої у будь-який момент може бути отримано інформацію для аналізу та побудови моделей. Дана інформація для побудови моделі має стосуватися лише тих кредитів, за якими вже відомо, чи були вони повернуті, чи ні. Вибірки слід обирати таким чином, щоб це була найновіша інформація щодо кредитів за один і той самий проміжок часу за одних і тих же умов. Тобто, перший-третій етапи організації аналізу даних — це фактично організація збору, перевірки та надходження в централізовану базу даних інформації щодо позичальників із відділень до центрального офісу.

Саме на четвертому етапі отримані статистичні дані завантажуються в блок аналізу даних, де на основі відомого методу будується модель. У цій інформаційній технології пропонується використати інтегрований метод і побудувати модель аналізу даних на основі дерева рішень та мереж Байєса. Спочатку на основі навчальної вибірки будується дерево рішень, яке визначає найсуттєвіші характеристики клієнта, що безпосередньо впливають на повернення кредиту. Далі в блок побудови мережі Байєса можна завантажити або лише характеристики, вибрані за допомогою дерева рішень (тобто, завантажити текстовий файл лише з цими даними), або на етапі побудови

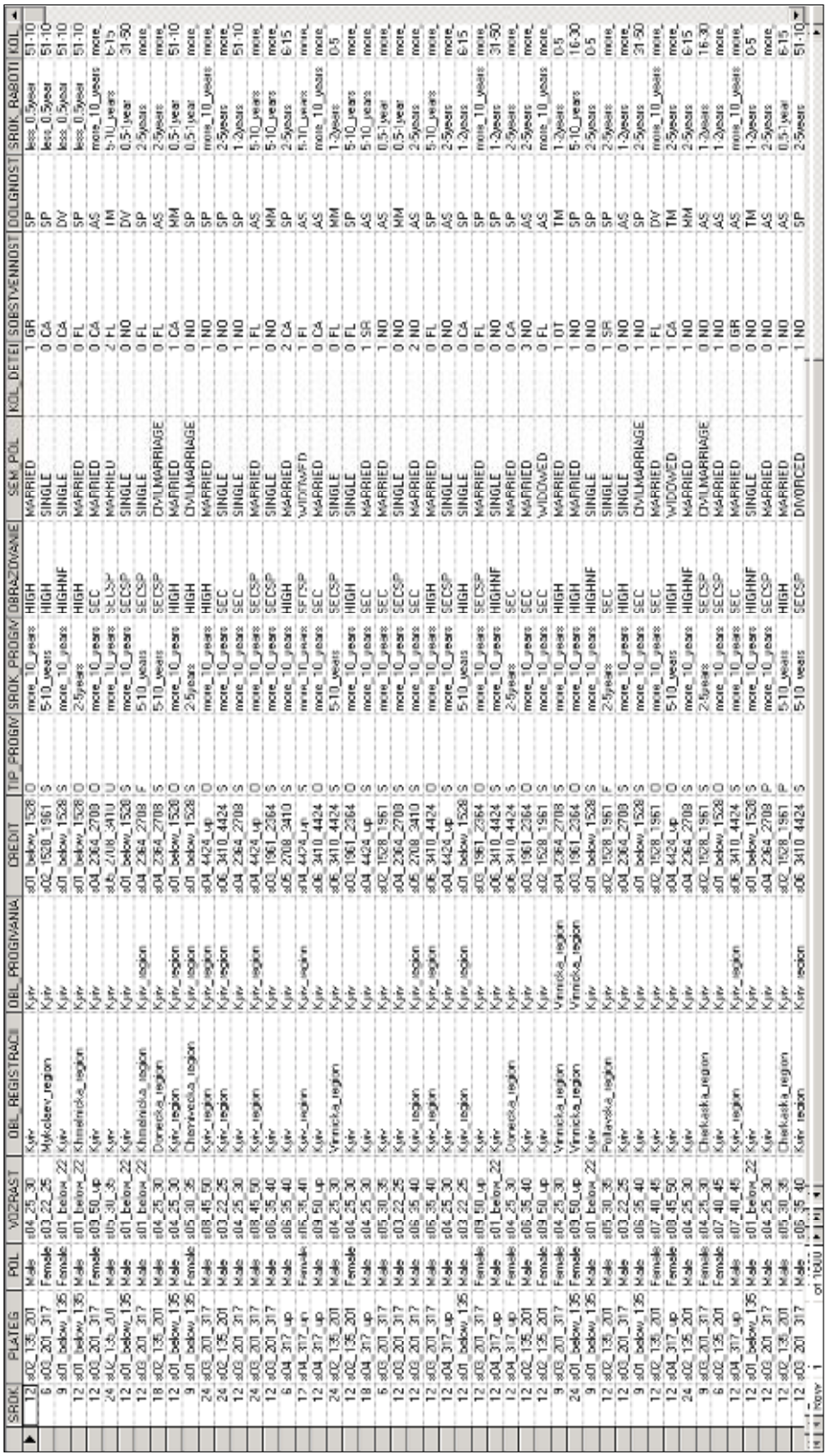
мережі вказати для побудови лише вибрані характеристики. Далі будується і навчається мережа Байєса. Отримана структура мережі використовується для аналізу характеристик моделі, побудованої на основі інтегрованого методу та оцінки кредитоспроможності позичальника. Для оцінювання характеристик моделей використовується перевірочна вибірка, для якої обчислюються ймовірності повернення кредиту. Отримані дані заносяться у спеціальний модуль, в якому автоматично обчислюються загальна точність, помилки 1-го та 2-го роду для різних порогів відсікання та будується ROC-крива [2, 5, 6]. На основі побудованої ROC-кривої обчислюється індекс GINI та визначається якість моделі (у випадку декількох моделей визначається краща модель).

ПРИКЛАД ЗАСТОСУВАННЯ ІНТЕГРОВАНОГО МЕТОДУ ДЛЯ АНАЛІЗУ КРЕДИТОСПРОМОЖНОСТІ ПОЗИЧАЛЬНИКА

Розглянемо окремі блоки інформаційної технології побудови та перевірки моделі на основі інтегрованого підходу. Для побудови моделі використовується статистика з 2200 випадків видачі кредитів, строк яких закінчився. Вибірка поділена на навчальну (2000 випадків) та перевірочну (200 випадків), вигляд якої показаний на рис. 3. Таким чином статистика зібрана і ми переходимо на другий крок алгоритму інтегрованого методу. Формалізуємо дані у зручному для застосування вигляді, тобто переводимо їх у заданий формат. Далі будуємо дерево рішень за допомогою одного з відомих програмних модулів. Змінна, що відображає інформацію, чи був кредит повернений — це залежна змінна, яка прогнозується, а характеристики клієнту та кредиту — це незалежні змінні.

За допомогою дерева рішень встановлено, що найсуттєвішими змінними, які впливають на змінну повернення кредиту та мають бути включені у модель на третьому кроці алгоритму є: ціна та тип товару, на який береться кредит; сімейний стан, вік, стать, освіта позичальника; кількість дітей в його сім'ї; стаж роботи та термін роботи на даному місці. На третьому кроці вибірка завантажується в модуль побудови мережі Байєса. Для побудови мережі використовуються змінні, відібрані на попередньому кроці. У результаті отримується структура мережі зображена на рис. 4.

Далі необхідно проаналізувати результати та перевірити якість моделі. Для цього визначається загальна точність, помилки першого та другого роду для різних порогів відсікання, будується ROC-крива та обчислюється індекс *GINI*. Для перевірки якості моделі використовується перевірочна вибірка, пороги відсікання встановлюються на рівні 0,9; 0,85; 0,8; 0,75 та 0,7. Інтегрована модель на основі дерева рішень та мережі Байєса проілюструвала прийнятне виявлення неплатоспроможних клієнтів у випадку консервативної політики банку. Кількість помилок першого роду, тобто пропуск дефолтів, становить 5%, що є кращим результатом порівняно з використанням методу дерев рішень. Необхідно зазначити, що модель забезпечує перестраховку, а тому буде корисно використовувати її тим банкам, які проводять консервативну політику та відсікають клієнтів із імовірністю повернення кредиту нижчою за 0,85–0,9. Для цієї моделі значення площі під ROC-кривою становить: $AUC = 0{,}784$, а індекс *GINI* віповідно: $GINI = 2 \times AUC - 1 = 0{,}568.$ Це прийнятні результати при оцінюванні якості моделі.



SROK	PLATEG	POL	VOZRAST	OBL_REGISTRACII	OBL_PROJIVANIA	CHEDIT	TIP_PROGM	SROK_PROGM	OBRAZOVANIE	SEM_POL	KOL_DETEI	SOBSTVENNOST	DOLGNOST	SROK_RABOTI	KOL
12	s02_135_201	Male	s04_25_30	Kyiv	Kyiv	s01_below_1528	O	more_10_years	HIGH	MARRIED	1	GR	SP	less_0.5year	51-10
6	s03_201_317	Female	s03_22_25	Mykolaev_region	Kyiv	s02_1528_1961	S	5-10_years	HIGH	SINGLE	0	CA	SP	less_0.5year	51-10
9	s01_below_135	Female	s01_below_22	Kyiv	Kyiv	s01_below_1528	S	more_10_years	HIGHNF	SINGLE	0	CA	DV	less_0.5year	51-10
12	s01_below_135	Male	s01_below_22	Khmelnicka_region	Kyiv	s01_below_1528	O	2-5years	HIGH	MARRIED	0	FL	SP	less_0.5year	51-10
12	s03_201_317	Female	s03_50_up	Kyiv	Kyiv	s04_2364_2708	O	more_10_years	SEC	MARRIED	0	CA	AS	more_10_years	more_
24	s4c_135_201	Male	s03_30_35	Kyiv	Kyiv	s4c_2708_3410	U	more_10_years	SECSP	MARRIED	2	FL	IM	5-10_years	5-15
12	s01_below_135	Male	s01_below_22	Kyiv	Kyiv	s01_below_1528	S	more_10_years	SECSP	SINGLE	0	NO	DV	0.5-1year	31-50
12	s03_201_317	Male	s01_below_22	Khmelnicka_region	Kyiv_region	s04_2364_2708	F	5-10_years	SECSP	SINGLE	0	FL	SP	2-5years	more_
18	s02_135_201	Male	s04_25_30	Donetska_region	Kyiv_region	s04_2364_2708	S	5-10_years	SECSP	CIVILMARRIAGE	0	FL	AS	2-5years	more_
12	s01_below_135	Male	s04_25_30	Kyiv_region	Kyiv_region	s01_below_1528	O	more_10_years	HIGH	MARRIED	1	CA	MM	0.5-1year	51-10
9	s01_below_135	Female	s05_30_35	Cherkaseka_region	Kyiv_region	s01_below_1528	S	2-5years	HIGH	CIVILMARRIAGE	0	NO	SP	0.5-1year	more_
24	s03_201_317	Male	s09_45_50	Kyiv_region	Kyiv_region	s04_4424_up	O	more_10_years	HIGH	MARRIED	1	NO	SP	more_10_years	more_
24	s02_135_201	Male	s03_22_25	Kyiv_region	Kyiv_region	s06_3410_4424	S	more_10_years	SEC	SINGLE	0	NO	SP	2-5years	more_
12	s03_201_317	Male	s04_25_30	Kyiv	Kyiv	s04_2364_2708	S	more_10_years	SEC	SINGLE	1	NO	SP	1-2years	51-10
24	s03_201_317	Male	s08_45_50	Kyiv_region	Kyiv_region	s04_4424_up	O	more_10_years	SECSP	MARRIED	1	FL	AS	5-10_years	more_
12	s03_201_317	Male	s06_35_40	Kyiv	Kyiv	s03_1961_2364	S	more_10_years	SECSP	SINGLE	0	NO	MM	5-10_years	more_
6	s04_317_up	Male	s06_35_40	Kyiv	Kyiv	s05_2708_3410	S	more_10_years	HIGH	MARRIED	2	CA	SP	2-5years	6-15
17	s04_317_an	Female	s05_30_40	Kyiv_region	Kyiv_region	s04_4424_up	S	more_10_years	SECSP	WIDOWED	1	FI	AS	5-10_years	more_
12	s04_317_up	Male	s09_50_up	Kyiv	Kyiv	s06_3410_4424	O	more_10_years	SEC	MARRIED	0	CA	AS	more_10_years	more_
24	s02_135_201	Male	s04_25_30	Vinnicka_region	Kyiv	s06_3410_4424	S	5-10_years	SECSP	SINGLE	0	FL	MM	1-2years	0-5
12	s02_135_201	Female	s04_25_30	Kyiv	Kyiv	s03_1961_2364	S	more_10_years	HIGH	SINGLE	0	FL	SP	5-10_years	more_
18	s04_317_up	Male	s04_25_30	Kyiv	Kyiv	s04_4424_up	S	more_10_years	SEC	MARRIED	1	SR	AS	5-10_years	more_
6	s03_201_317	Male	s05_30_35	Kyiv	Kyiv	s02_1528_1961	S	more_10_years	SEC	MARRIED	1	NO	AS	0.5-1year	more_
12	s03_201_317	Male	s03_22_25	Kyiv	Kyiv	s04_2364_2708	S	more_10_years	SECSP	SINGLE	0	NO	MM	0.5-1year	more_
12	s03_201_317	Male	s06_35_40	Kyiv_region	Kyiv_region	s05_2708_3410	S	more_10_years	SEC	MARRIED	2	NO	AS	2-5years	more_
12	s03_201_317	Male	s05_35_40	Kyiv_region	Kyiv_region	s06_3410_4424	O	more_10_years	HIGH	MARRIED	0	FL	SP	5-10_years	more_
12	s04_317_up	Male	s04_25_30	Kyiv	Kyiv	s04_4424_up	S	more_10_years	SECSP	SINGLE	0	NO	AS	2-5years	more_
12	s01_below_135	Male	s03_22_25	Kyiv_region	Kyiv_region	s01_below_1528	S	5-10_years	HIGH	SINGLE	0	CA	SP	1-2years	6-15
12	s03_201_317	Female	s09_50_up	Kyiv	Kyiv	s03_1961_2364	O	more_10_years	SECSP	MARRIED	0	FL	SP	more_10_years	more_
12	s04_317_up	Male	s01_below_22	Kyiv	Kyiv	s06_3410_4424	S	more_10_years	HIGHNF	SINGLE	0	NO	SP	1-2years	31-50
12	s04_317_up	Female	s04_25_30	Donetska_region	Kyiv	s06_3410_4424	S	2-5years	SEC	MARRIED	0	CA	AS	2-5years	more_
12	s02_135_201	Male	s06_35_40	Kyiv	Kyiv	s02_1961_2364	O	more_10_years	SEC	MARRIED	3	NO	AS	2-5years	more_
12	s02_135_201	Male	s09_50_up	Kyiv	Kyiv	s02_1528_1961	S	more_10_years	SEC	WIDOWED	0	FL	AS	more_10_years	more_
9	s03_201_317	Female	s04_25_30	Vinnicka_region	Vinnicka_region	s04_2364_2708	S	more_10_years	HIGH	MARRIED	1	OT	TM	1-2years	0-5
24	s01_below_135	Female	s09_50_up	Vinnicka_region	Vinnicka_region	s01_1961_2364	O	more_10_years	HIGH	MARRIED	1	NO	SP	5-10_years	16-30
9	s01_below_135	Male	s01_below_22	Kyiv	Kyiv	s01_below_1528	S	more_10_years	HIGHNF	SINGLE	0	NO	SP	2-5years	0-5
12	s02_135_201	Male	s05_30_35	Poltavska_region	Kyiv	s02_1528_1961	F	2-5years	SEC	SINGLE	1	SR	SP	2-5years	more_
12	s03_201_317	Male	s03_22_25	Kyiv	Kyiv	s04_2364_2708	S	more_10_years	HIGH	SINGLE	0	NO	AS	1-2years	more_
9	s01_below_135	Male	s06_35_40	Kyiv	Kyiv	s01_below_1528	S	more_10_years	SEC	CIVILMARRIAGE	1	NO	SP	2-5years	31-50
12	s03_201_317	Female	s07_40_45	Kyiv	Kyiv	s02_1528_1961	O	more_10_years	SEC	MARRIED	1	FL	DV	more_10_years	more_
12	s04_317_up	Female	s09_45_50	Kyiv	Kyiv	s04_4424_up	O	5-10_years	HIGH	WIDOWED	1	CA	TM	2-5years	more_
24	s02_135_201	Male	s04_25_30	Kyiv	Kyiv	s04_2364_2708	S	more_10_years	HIGHNF	MARRIED	1	NO	MM	2-5years	6-15
9	s03_201_317	Female	s04_25_30	Cherkaseka_region	Kyiv	s02_1528_1961	S	2-5years	SECSP	CIVILMARRIAGE	0	NO	AS	1-2years	16-30
6	s02_135_201	Female	s07_40_45	Kyiv	Kyiv	s01_below_1528	O	more_10_years	SECSP	MARRIED	1	NO	AS	1-2years	more_
12	s04_317_up	Male	s07_40_45	Kyiv_region	Kyiv_region	s06_3410_4424	S	more_10_years	SEC	MARRIED	0	GR	AS	more_10_years	more_
12	s01_below_135	Male	s01_below_22	Kyiv	Kyiv	s01_below_1528	S	more_10_years	HIGHNF	SINGLE	0	NO	TM	1-2years	0-5
12	s03_201_317	Male	s04_25_30	Kyiv	Kyiv	s04_2364_2708	P	more_10_years	SECSP	MARRIED	0	NO	AS	2-5years	more_
12	s02_135_201	Male	s05_30_35	Cherkaseka_region	Kyiv	s02_1528_1961	P	5-10_years	HIGH	MARRIED	1	NO	AS	0.5-1year	6-15
12	s03_201_317	Male	s06_35_40	Kyiv_region	Kyiv_region	s06_3410_4424	S	5-10_years	SECSP	DIVORCED	1	NO	SP	2-5years	51-10

Рис. 3. Приклад завантаженої навчальної вибірки

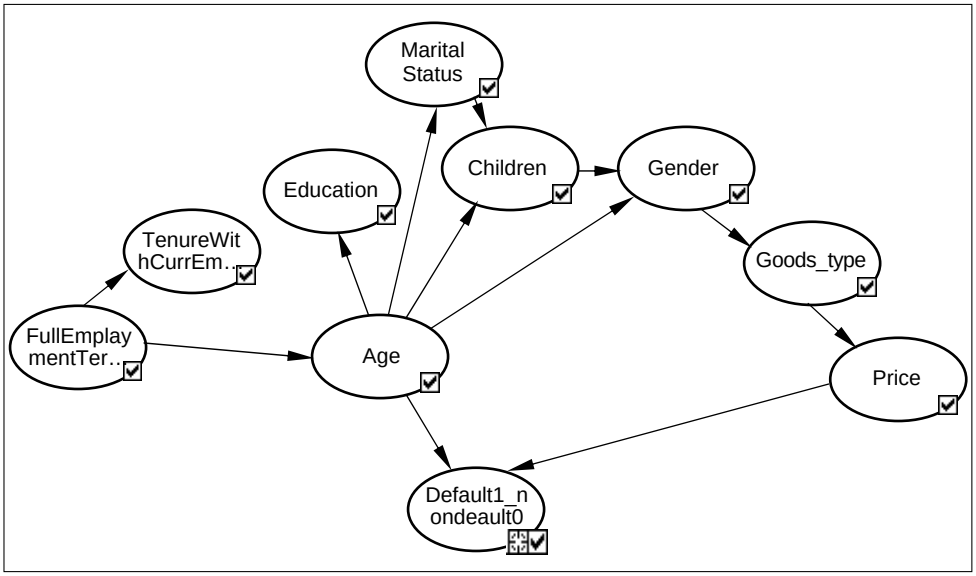


Рис. 4. Приклад структури мережі Байєса, побудованої на основі інтегрованого методу

Отриману та протестовану модель у вигляді програмного модуля встановлюють на місцях працівникам-менеджерам банку, які видають кредити для того, щоб вони швидше отримували інформацію щодо можливості видачі кредиту. Якщо система видає інформацію, що кредит можна видати, то менеджери банку передають усі дані позичальника та інформацію по кредиту в базу даних або інформаційну систему, обов'язково визначаючи, яку ймовірність повернення кредиту видала модель. Зважаючи на гнучку політику банку щодо кредитування (періодичне підвищення або пониження порогу відсікання), є сенс надати можливість змінювати поріг відсікання клієнтів у процесі функціонування цієї моделі. Для цього в разі необхідності можна вивантажити дані по кредитах, які було видано з початку застосування цієї моделі, за якими було встановлено ймовірності їх повернення, і за якими вже наявна інформація про повернення кредиту. Для банків, які проводять консервативну політику видачі кредитів, такий поріг встановлюється вищим, щоб відсіяти якомога більше нестабільних та ненадійних клієнтів, для яких ймовірність повернення кредиту нижча за обраний поріг відсікання. У разі проведення банками агресивної політики кредитування, тобто, коли банк хоче завоювати велику кількість нових клієнтів він навмисне опускає поріг відсікання, щоб видати якомога більше кредитів, заробивши при цьому великий прибуток. Таку політику проводили дочірні банки великих банків із іноземним капіталом, видаючи кредити, вже за перший рік кредитування, виправдовуючи закладені ресурси. Скоріш за все, саме така агресивна політика кредитування буде спостерігатися найближчим часом, коли банки вирішать масово відновлювати усі види кредитування (споживче, іпотечне тощо) і той, хто перший розпочне цей процес, буде встановлювати вигідні для себе умови, заробляючи при цьому надвеликі прибутки.

ВИСНОВКИ

У роботі запропоновано новий інтегрований метод аналізу даних, перевагою якого є те, що він дозволяє обробляти дані та встановлювати взаємозалежності між змінними там, де інші методи не можуть бути застосовані без втрати певної інформації. Запропонована інформаційна технологія аналізу даних позичальників на основі інтегрованого методу дозволяє побудувати адекватну модель позичальника, оцінити кредитоспроможність та спрогнозувати ймовірність повернення кредиту. Ця технологія використовувалась разом із відомими підходами до оцінювання позичальників та дозволила отримати додаткову оцінку під час прийняття рішень щодо видачі кредиту, але не викликала жодних проблем із навчанням персоналу. Використання запропонованої інформаційної технології дає можливість скоротити обсяги можливих втрат від несумлінних позичальників завдяки оптимізації розрахункових операцій, прогнозування надходжень та витрат і планування розподілу коштів. Інтегрований метод апробовано на фактичних прикладах оцінювання даних позичальників, показав прийнятні за точністю результати, а тому його можна застосовувати до аналізу інших типів фінансових даних та прикладних областей.

У подальших дослідженнях планується вдосконалити структуру запропонованого інтегрованого методу, розширити його функціональні можливості та автоматизувати окремі етапи обробки даних. Все це сприятиме підвищенню якості обробки даних та скороченню часу на прийняття обґрунтованих об'єктивних рішень щодо видачі кредитів.

ЛІТЕРАТУРА

1. *Kiss F.* Credit scoring processes from a knowledge management perspective // Periodica Polytechnica Series: Society, Man, Sybernetics. — 2003. — **11**. — № 1. — P. 95–110.
2. *Кузнєцова Н.В., Бідюк П.І.* Порівняльний аналіз характеристик моделей оцінювання ризиків кредитування // Наукові вісті НТУУ «КПІ». — 2010. — № 1. — С. 42–53.
3. *Терентьєв А.Н., Бидюк А.В., Миронова А.В.* и др. Сравнение методов интеллектуального анализа данных при оценивании кредитоспособности физических лиц // Проблемы управления и информатики. — 2009. — № 5. — С. 141–149.
4. *Кузнєцова Н.В., Бідюк П.І.* Системний підхід до аналізу кредитних ризиків з використанням мереж Байєса // Наукові вісті НТУУ «КПІ». — 2008. — № 3. — С. 11–24.
5. *Кузнєцова Н.В.* Методи оцінювання ризиків роздрібного кредитування // Системний аналіз та інформаційні технології: матеріали XII міжнар. наук.-техн. конф. SAIT–2010, 25–29 травня 2010 р.: тези доп. — Київ: ННК «ІПСА» НТУУ «КПІ», 2010. — С. 272.
6. *Кузнєцова Н.В.* Інтегрований підхід до оцінювання кредитних ризиків // Труды Одесского политехн. ун-та. — 2010. — №1 (33). — С. 157–165.

Надійшла 04.06.2010