УДК 62-50

ОСНОВНІ ЕТАПИ ПОБУДОВИ І ПРИКЛАДИ ЗАСТОСУВАННЯ МЕРЕЖ БАЙЄСА

П.І. БІДЮК, Н.В. КУЗНЄЦОВА

Розглянуто основні формалізми мереж Байєса (МБ), їх графотеоретичне представлення. Проаналізовано методи побудови та навчання таких мереж. На основі аналізу запропоновано та обґрунтовано методику побудови і застосування МБ для конкретних задач, яка апробована при розв'язанні економічних та фінансових задач (кредитування, аналізу причин повернення товарів) з використанням різних програмних засобів.

ВСТУП

Розвиток систем підтримки прийняття рішень (СППР) і впровадження їх у багатьох сферах діяльності постійно вимагає розробки нових математичних, технологічних і програмних засобів. Одним із актуальних питань є прийняття рішень з урахуванням невизначеності.

Відомо, що більшість з систем прийняття рішень у реальному часі базується на правилах. Введення у такі системи ймовірнісних оцінок та методів відображення причинно-наслідкової залежності дозволяє розширити область застосування СППР та покращити якість запропонованих рішень. Математичний апарат мереж Байєса (МБ) дозволяє поєднати досить просте графічне представлення деякого процесу з його ймовірнісним характером, проаналізувати можливі варіанти розвитку ситуації, відслідкувати правильність встановлення причинно-наслідкового зв'язку між окремими подіями і завдяки цьому підвищити обґрунтованість рішень при аналізі складних проблемних ситуацій.

Мета даного дослідження — розробка практичної методики побудови МБ на основі статистичних даних, що описують модельований процес, та ілюстрація можливостей їх застосування.

Побудова МБ

Одним з основних етапів побудови МБ для вирішення конкретної прикладної задачі є формування графо-теоретичної моделі.

Нехай $G = (V, Bi)$ — деякий граф, в якому V — скінченна множина змінних; Bi — нерефлексивне бінарне відношення на V. Якщо дуги цього графу спрямовані, то його називають *спрямованим ациклічним графом* (САГ). Якщо спрямований граф містить хоча б один замкнений цикл, то його називають *спрямованим циклічним графом* (СЦГ).

Деякі змінні МБ мають спеціальні імена. Змінну (вузол) називають *нащадком*, якщо вона залежить від однієї або більше інших змінних. *Батьківська* змінна має одну або більше змінних-нащадків. До множини *нащадків* відносять змінні-нащадки, які є нащадками однієї і тієї ж батьківської змін-

© *П.І. Бідюк, Н.В. Кузнєцова*, 2007

ної, а також змінні-нащадки змінних-нащадків вищого рівня. Одна і та ж змінна може бути батьківською змінною і нащадком одночасно. Якщо змінна не має жодної батьківської, то її називають *кореневою*. Кореневі змінні — найважливіші змінні мережі. Це саме те, що нас цікавить. Змінну називають *листком*, якщо вона не має змінних-нащадків.

Формально МБ — це трійка $N = <V, G, J>$, першою компонентою якої є множина змінних V; другою — САГ G, вузли якого відповідають випадковим змінним модельованого процесу; 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, за винятком батьківських попередників цієї змінної.

Визначення апріорних ймовірностей подій і формула Байєса

Розглянемо приклад мережі [4], у якій ймовірність перебування вершини e у різних станах (e_k) залежить від станів (c_i, d_j) вершин c та d і визначається виразом $p(e_k) = \sum_i \sum_j p(e_k \mid c_i, d_j) \times p(c_i, d_j)$, де $p(e_k \mid c_i, d_j)$ — ймовірність перебування в стані e_k у залежності від станів c_i, d_j (рис.1). Оскільки події, що представлені вершинами c і d незалежні, то $p(e_k \mid c_i, d_j) = p(c_i) p(d_j)$.

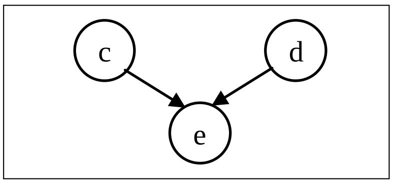

Рис. 1. Приклад найпростішої мережі довіри Байєса

Розглянемо приклад більш складної мережі (рис. 2).

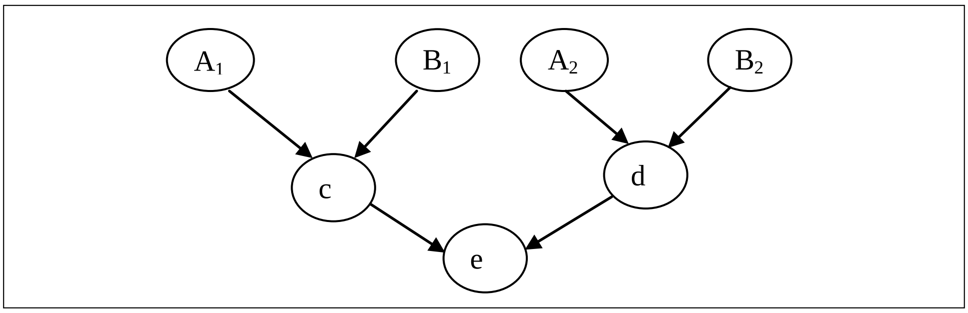

Рис. 2. Дворівнева мережа довіри Байєса

Рис. 2 ілюструє умовну незалежність подій. Для оцінки вершин c і d використовуються ті самі вирази, що й для обчислення ймовірності $p(e_k)$, тоді

$$p(c_i) = \sum_m \sum_n p(c_i \mid A_{1m}, B_{1n}) \times p(A_{1m}) \times p(B_{1n}),$$

$$p(d_j) = \sum_m \sum_n p(d_j \mid A_{2m}, B_{2n}) \times p(A_{2m}) \times p(B_{2n}).$$

З цих виразів видно, що вершина e умовно не залежить від вершин A_1, A_2, B_1, B_2, оскільки не має стрілок, що безпосередньо поєднують ці вершини. Саме процес обчислення ймовірностей стає основою для формування рішень в умовах невизначеності на основі МБ. Основу байєсівського підходу складає поняття умовної ймовірності $P(A|B) = x$. Це означає, що при умові виникнення B (і всього іншого, що не має відношення до B), ймовірність виникнення A дорівнює x. Спільна ймовірність появи подій A і B визначається за формулою повної ймовірності $P(A,B) = = P(A|B)P(B) = P(B|A)P(A)$.

Це рівняння є *фундаментальним правилом* для обчислення ймовірностей і основою для *теореми Байєса* $P(B|A) = \dfrac{P(A|B)P(B)}{P(A)}$. З формулою повної ймовірності тісно пов'язана формула Байєса. Якщо до виконання експерименту ймовірності гіпотез були $P(H_1), P(H_2), ..., P(H_n)$, а в результаті експерименту з'явилась подія A, то з урахуванням цієї події «нові», тобто умовні ймовірності гіпотез, обчислюються за формулою Байєса

$$P(H_k / A) = \frac{P(H_k)P(A/H_k)}{P(A)} (k = 1,2,...,n),$$

де $P(A) = \sum_{i=1}^{n} P(H_i)P(A/H_i)$.

Умовна ймовірність $P(H_k / A)$ може бути визначена як відношення ваги дуги, що проходить через вершину, яка відповідає гіпотезі H_k, до ваги всього ймовірнісного графа. Теорема Байєса застосовується у випадках, коли існує інформація про залежні змінні (*свідоцтва*), а суть дослідження полягає у визначенні ймовірностей вихідних змінних (*причин*). Так, за умови наявності умовної ймовірності $P(B|A)$ виникнення деякої події B при умові, що має місце подія A, теорема Байєса дає розв'язок оберненої задачі, якою є ймовірність виникнення події A, якщо подія B відбулася.

Для МБ введені такі базові поняття.

- *Ланцюгове правило* — це засіб обчислення повної ймовірності у байєсових мережах. Якщо МБ визначена на множині вершин $U = = \{A_1, A_2, ..., A_n\}$, то спільний розподіл ймовірностей є добутком усіх умовних ймовірностей, визначених у МБ.

$$P(U) = \prod_i P(A_i | par(A_i)),$$

де $par(A_i)$ — множина (станів) батьківських вершин для A_i.

- *Умовна незалежність* вершин МБ — блокування впливу між цими вершинами. Змінні (множини змінних) A і C є незалежними при відомому стані змінної B, якщо $P(A|B) = P(A|B,C)$. Коли стан вершини B відомий, то ніяка інформація про C не змінює ймовірностей A.

- *Маржиналізація* — підсумовування ймовірностей на реалізаціях усіх змінних, окрім вибраних. Вона використовується для обчислення ймовірностей змінних, які нас цікавлять, на основі повної ймовірності $P(A) = \sum_{B} P(A, B)$.

- *Ймовірність Байєса* для деякої події x — міра упевненості деякого експерта, що дана подія відбудеться [3]. У той час як класична ймовірність — це фізична властивість реального світу, ймовірність Байєса — це більше атрибут експерта, що її визначає.
- *Логічний висновок* у мережах Байєса — обчислення умовних ймовірностей деяких змінних на основі наданої інформації про інші змінні.
- *Розповсюдження* в МБ — процес обчислення апостеріорних ймовірностей для тих вузлів мережі, які не спостерігаються, на основі значень спостережуваних вузлів (тобто свідоцтв).

Логічний висновок у МБ

Ключовим поняттям обчислення ймовірностей у МБ є *процес оновлення ймовірностей*, або зміна міри довіри. Алгоритм цього процесу визначає спосіб отримання апостеріорних ймовірностей вершин мережі на основі отриманої інформації. Оновлення міри довіри вершин може розглядатися як синонім логічного висновку в МБ, а довіра по відношенню до них — це розподіл ймовірностей, обумовлений усіма свідоцтвами, що поступили у мережу.

Нехай МБ є мережею на множині змінних U, і нехай e — множина тверджень вигляду *змінна* A *знаходиться у стані* a. Таким чином, e представляє собою твердження *спільна конфігурація вершин* $A,\ldots,B$ *задана як* $a,\ldots,b$. При розв'язанні задач ми прагнемо знайти апостеріорний розподіл ймовірностей $P(X \mid e)$ для усіх змінних $X \in U$.

Математично ця задача може бути розв'язана наступним чином:

- використати ланцюгове правило для обчислення $P(U)$;
- відокремити $P(U, e)$ — частину $P(U)$, яка відповідає конфігурації $(a,\ldots,b)$;
- отримати $P(X, e)$ шляхом маржиналізації $P(U, e)$ для кожного $X \in U$ (тобто для кожного стану $x \in X$ підсумувати усі елементи $P(U, e)$, для яких X знаходиться у стані x);
- обчислити $P(X \mid e)$ як результат нормалізації $P(X \mid e)$, тобто розділити $P(X \mid e)$ на суму всіх його членів.

Однак зазвичай $P(U)$ настільки об'ємна, що її неможливо зберігати, або необхідні обчислення можуть виявитися неприйнятно об'ємними. Відмовитися від використання повної ймовірності $P(U)$ вдається при послідовному застосуванні теореми Байєса.

Дійсно, задача знаходження ймовірностей $P(X \mid e)$ на основі множини свідоцтв e відносно МБ може бути представлена як задача оновлення ймовірностей на (під)мережі, до складу якої входить лише певна підмножина вузлів графу. Зменшення мережі до цієї підмножини відбувається шляхом послідовного виключення вузлів (маржиналізації) і механізму інверсії ребер графу на основі теореми Байєса. На рис. 3 показано отримання розподілу $P(C)$ шляхом послідовної редукції графа виключенням вершин (кроки 1, 2, 4, 5) та інверсії ребер (крок 3).

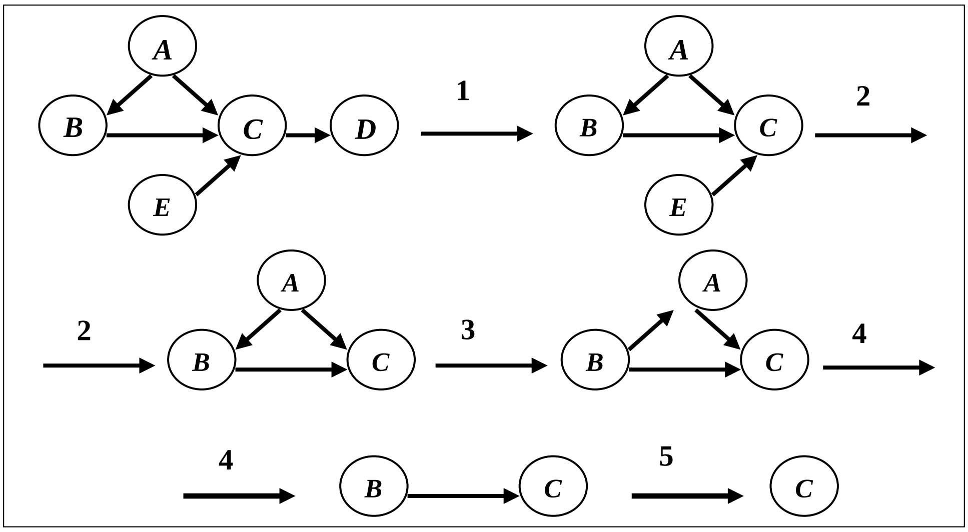

Рис. 3. Приклад ймовірнісного висновку

Необхідно зазначити: наслідком теореми Байєса є підтримка нею оцінювання ймовірностей на графі в обох напрямах. Процес формування висновку в мережі супроводжується розповсюдженням по мережі свідчень, що надійшли. При цьому процес розповсюдження ймовірностей у МБ грунтується на механізмі їх перерахунку [4]. Для спрощення обчислювального процесу використовується метод зашумленого вентиля АБО (noisy or gate) [5, 6]. Його суть полягає у тому, що в деяких прикладах вершина y може бути умовно незалежною від цілого ряду вершин x_r, де $r = 1, 2, ..., n$. Цей метод використовують для обчислення 2^n ймовірностей, які необхідні при використанні таблиць умовних ймовірностей. Згідно із цим методом ймовірність y в залежності від n вершин x_r оцінюється як $p(y \mid x_1, x_2, ..., x_n) = = 1 - \prod_{i=1}^{n}(1 - p(y \mid x_i))$, що дозволяє оцінити лише тільки ймовірності $p(y \mid x_1), p(y \mid x_2), ..., p(y \mid x_n)$ і на їх основі визначити $p(y \mid x_1 x_2, ..., x_n)$.

Слід зазначити, що для визначення наступного вузла, який виключається, необхідно розглянути усю мережу повністю. Вплив свідоцтва простежується лише на окремий вузол, а для виключених вузлів він невідомий. І, нарешті, наведений алгоритм концептуально послідовний, хоча для побудови життєздатних моделей людських міркувань паралелізм видається більш адекватним.

У 1986 р. опубліковано *алгоритм формування логічного висновку* в МБ, що дозволяє уникнути цих недоліків [9]. Його сутність полягає у використанні поняття *повідомлення*, згідно з яким оновлення ймовірностей вершин мережі здійснюється шляхом розсилання кожною вершиною мережі двох типів повідомлень про свій стан:

1) батьківським вершинам — π -повідомлення;

2) дочірнім — λ -повідомлення.

Міра довіри для події $X = x$ розраховується як нормований добуток числового еквіваленту повідомлень $\lambda(x)$ і $\pi(x)$. Не наводячи повного математичного опису цього алгоритму, розглянемо деякі аспекти процесу обміну повідомленнями для обчислення апостеріорних ймовірностей.

Якщо отримано новий розподіл $P^*(A)$ для вершини A, то для обчислення нового B розподілу $P^*(B) = \sum_A P(B|A)P^*(A)$ можна скористатися фундаментальним правилом обчислення ймовірностей і маржиналізацією.

Розповсюдження в напрямі зв'язків досить просте: це повідомлення, надіслане від A до B. Це ж повідомлення і є розподілом A, і на цій основі відбувається оновлення розподілу B. Оновлення довіри відбувається і в протилежному напрямі: інформація щодо B може бути використана для зміни міри довіри A. Інструментом розповсюдження ймовірностей у зворотному напрямі є теорема Байєса.

Розглянемо МБ, граф якої є *деревом*. Повідомлення можуть надсилатися в обох напрямах, тобто вершина дерева X може надіслати повідомлення до будь-якої сусідньої вершини Y. Повідомлення, що надсилається, є поточним розподілом X, і вершина Y використовує це повідомлення для оновлення власного розподілу.

Якщо вершини надсилають повідомлення неупорядковано, переходячи у деякий момент часу в режим очікування, то можна довести, що у такому разі існує стан рівноваги (*стійкий стан*), коли жодне подальше повідомлення не змінює жодного розподілу. Усталений стан досягається після скінченного числа пересилань, і у цьому стані кожна вершина зберігає коректний розподіл ймовірностей.

Пересилання повідомлень упорядковується за допомогою такого правила: *вершина X може надіслати повідомлення своєму сусіду Y*, якщо X отримала повідомлення від усіх своїх інших сусідів. У цьому разі листки дерева розпочинають надсилати повідомлення і хоча б одна вершина буде мати можливість надсилання повідомлень до тих пір, поки повідомлення не буде надіслано по кожному ребру дерева уздовж обох напрямків. Якщо це сталося, то дерево знаходиться у стійкому стані.

Отже, в алгоритмі, що грунтується на повідомленнях, вплив кожного нового свідоцтва розглядається як збурення, що розповсюджується по мережі шляхом обміну повідомленнями між сусідніми вершинами. Досягнення стійкого стану при застосуванні цього алгоритму, як і гарантія коректності результатів, вимагають існування деревоподібної структури мережі. МБ має бути полідеревом, тобто САГ, в якому між будь-якими двома вершинами існує лише один маршрут без урахування напрямку ребер. Це суттєве обмеження, оскільки більшість реальних предметних процесів неможливо змоделювати на графі без циклів.

Для того щоб обійти це обмеження запропоновано декілька методів. Так, *метод обумовлення* враховує, що будь-яка мережа з циклом може бути зведена до полідерева перебором станів кореневої вершини циклу (послідовним обумовленням цієї вершини). Коли вершина обумовлена, тобто знаходиться у певному стані, можна вважати, що вона більше не є частиною мережі, а замість неї розглянути стільки мереж, скільки станів ця вершина має. Після цього теорема Байєса дозволить обчислити умовні ймовірності станів цієї вершини при відомих розподілах ймовірностей її сусідських (дочірніх) вершин. Необхідно обчислити стільки мереж, скільки станів має коренева вершина кожного циклу, а тому цей метод уповільнює обчислення.

Інша ідея використовується у *методі кластеризації* [9], розвиненому в роботах [7, 10]. У ролі вершин дерева у ньому виступають не окремі вершини, а їх групи (кластери). За допомогою методів теорії графів аналізуються властивості незалежності вершин мережі і формується множина кластерів, які поєднують у дерево. Отримане дерево має властивість *сполученого дерева*: для кожної пари (V,W) вершин дерева усі вершини маршруту між V та W містять їх перетин $V \bigcap W$. Сполучене дерево, отримане в результаті кластеризації вершин вихідного дерева, є полідеревом, і в ньому можна застосувати обмін повідомленнями. Після отримання апостеріорних ймовірностей для кластерів обчислюються ймовірності вершин вихідного дерева.

Метод кластеризації є одним з найпотужніших серед методів точного обчислення ймовірностей на МБ. Поява і швидкий розвиток наближених методів пояснюються тим, що в загальному випадку при зростанні кількості змінних точне обчислення ймовірностей є вкрай складним. Задача точного обчислення ймовірностей у МБ відноситься до класу NP-повних (причиною нелінійної поліноміальної складності обчислень найчастіше є топологія мережі, що може містити цикли). Проте встановлено [6, 7], що і наближений ймовірнісний висновок у МБ також є NP-складним. У практичних задачах широко використовуються саме наближені методи.

Структура мережі найчастіше визначається експертами предметної області, хоча існують методи структурного навчання МБ на основі даних. Таблиці умовних ймовірностей, навпаки, часто генеруються на основі даних за допомогою статистичних методів. Проте слід підкреслити, що принципово суб'єктивний байєсівський підхід не вимагає «об'єктивності» ймовірностей, а тому дозволяє при формуванні таблиць умовних ймовірностей спиратися на суб'єктивні оцінки експертів. Слід також зазначити, що результати логічного висновку більш чутливі до якісної структури МБ, ніж до кількісних значень ймовірностей [10].

Навчання структури МБ

Для опису МБ необхідно визначити дві характеристики: топологію графа (структуру) і параметри кожного умовного розподілу ймовірностей (УРЙ). Однак навчання структури набагато складніше, ніж навчання параметрів. У роботах [1,8] розглянуті основні чотири випадки, які мають місце при навчанні мережі.

1. Відома структура, повна спостережуваність.

Використовується метод максимальної оцінки правдоподібності за таблицею умовних ймовірностей.

2. Відома структура, часткова спостережуваність.

Використовується ЕМ-алгоритм (максимізації математичних сподівань) для знаходження (локально) оптимальної максимальної оцінки ймовірності параметрів.

3. Невідома структура, повна спостережуваність.

У роботі [8] для даного випадку визначається функція оцінювання, що використовується для вибору моделі, потім обговорюються алгоритми, які скеровані на оптимізацію цієї функції у просторі моделей, і, нарешті, досліджуються їх обчислювальна і типова складність.

4. Невідома структура, часткова спостережуваність.

Розглянемо докладніше цей найскладніший випадок: структура невідома і є приховані змінні і/або втрачені дані. Для того щоб обчислити ваговий коефіцієнт Байєса, потрібно розглянути другорядні приховані вузли і параметри. Оскільки це дуже складно, то зазвичай використовують асимптотичне наближення до наступного критерію, названого байєсівським інформаційним критерієм (БІК), який визначається таким чином:

$$\log(P(G|D)) \approx \log(P(D|G, \hat{\theta}_G)) - \frac{\log(N)}{2}\dim(G),$$

де N — кількість зразків; $\hat{\theta}_G$ — ML-оцінка параметрів; $\dim(G)$ — розмірність моделі. Перший вираз — це лише ймовірність, а другий — штраф за складність моделі. Хоча БІК розкладається у суму локальних виразів по одному до кожного вузла локальний пошук все ж таки складний, оскільки ми маємо виконати ЕМ-алгоритм на кожному кроці. Альтернативний підхід полягає у тому, щоб виконати кроки локального пошуку на M-му кроці ЕМ–алгоритму. Він називається структурованим, якщо сходиться до локального максимуму ваги БІК.

Приклади застосування МБ

Практично, на сьогоднішній день реалізовано не так вже й багато СППР, які включали б апарат МБ, однак їх суттєві переваги в умовах невизначеності привели до застосування МБ при вирішенні найбільш складних задач медицини, фізики, хімії, інтелектуальної обробки даних тощо. Розглянемо деякі цікаві підходи, використані при створенні відповідних СППР.

Найбільш широко вживані МБ [8] — це, без сумніву, вбудовані у продукти Microsoft, включаючи Answer Wizard of Office 95, Office Assistant of Office 97 і більш ніж 30 аварійних програм технічної підтримки.

МБ спочатку з’явилися як спроба додати ймовірнісні змінні до експертних систем, і це найбільша галузь їхнього вживання. Відомий приклад — QMR-DT стосується переформулювання моделі Швидкої Медичної Довідки (QMR) в теорії прийняття рішень [8].

Інше цікаве застосування МБ — система Vista, розроблена Е. Хорвітцом [8], яка використовувалась у Центрі управління польотами НАСА у Хьюстоні протягом декількох років. Система послуговується МБ для того, щоб відображати телеметрію у реальному часі, і забезпечувати довідку щодо відносної ймовірності альтернативних відмов космічних систем. Вона також враховує критичність часу і рекомендує дії найбільш очікуваної корисності, керує представленням інформації з метою динамічного визначення найбільш важливої інформації. Е. Хорвітц спробував використати подібну технологію в продуктах Microsoft, зокрема, у проекті Lumiere.

Однак ці приклади застосування МБ — це лише невелика частка того, що може бути реалізавано і вдало використано. Наприклад, доцільно спробувати застосувати цей потужний інструмент для розв'язання фінансово-економічних задач. Для цього спочатку розглянемо формалізовану методику побудови МБ.

Методика побудови і застосування МБ

При побудові МБ з метою вирішення конкретної задачі необхідно:

1. Виконати аналіз проблеми і дати формалізовану постановку задачі. Сформулювати питання, на яке має бути отримана ймовірнісна відповідь у результаті формування ймовірнісного висновку за допомогою побудованої мережі.

2. Визначити множину даних, що відносяться до змінних задачі, отримати їх експертні оцінки та/або статистичні дані.

3. Поставити у відповідність усім отриманим даним взаємовиключні змінні.

4. Побудувати ациклічний граф, що відображає істотні умови незалежності змінних та існування причинно-наслідкових зв'язків.

5. Визначити апріорні ймовірності та оптимізувати топологію мережі на основі наявної інформації.

6. Виконати навчання мережі і провести формування висновку по відношенню до відповідних станів процесу.

7. Обробити результати: проаналізувати їх і зробити висновки щодо ймовірності очікуваної події.

Розглянемо приклад реалізації сформульованої методики при вирішенні досить простої задачі аналізу причин повернення товарів.

Іноземна компанія, яка реалізує свою продукцію в Україні, збирає статистику щодо причин повернення товарів на склад. Розглядаються можливі варіанти: товар пошкоджений у процесі транспортування чи товар бракований. Із статистичних даних відомо, що ймовірність повернення товару через пошкодження при транспортуванні становить 0,05, а ймовірність повернення через брак — 0,3. Необхідно встановити ймовірність того, що повернений товар був бракований.

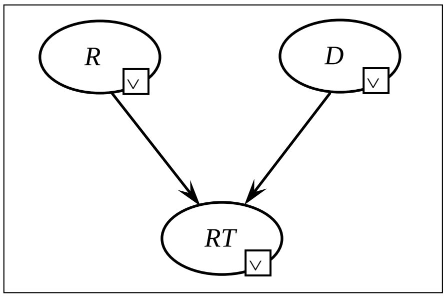

Рис. 4. Приклад мережі довіри Байєса для задачі повернення товарів

Використовуючи цю інформацію, будуємо мережу довіри Байєса, де R (Reject) — повернення товару саме через брак, D (Damaged) — повернення через пошкодження товару, RT (Return) — повернення товару на склад (рис. 4).

Ймовірність повернення товару для даної мережі обчислюється за формулою спільної ймовірності $P(RT) = P(R)\,P(D)\,P(RT \mid R, D)$. Значення ймовірностей вузлів R і D наведені у табл. 1, а значення умовних ймовірностей для вузла RT — у табл. 2.

Таблиця 1. Значення ймовірностей вузлів

Prob.	TRUE	FALSE
$P(D)$	0,05	0,95
$P(R)$	0,3	0,7

Таблиця 2. Значення умовних ймовірностей

$P(RT \mid R, D)$	$D = \text{TRUE}$	$D = \text{FALSE}$
$R = \text{TRUE}$	(0,9; 0,1)	(0,5; 0,5)
$R = \text{FALSE}$	(0,2; 0,8)	(0,1; 0,9)

Використовуючи теорему Байєса, можна обчислити ймовірність повернення товару через брак товару.

$$P(R = \text{TRUE} \mid RT = \text{TRUE}) = \frac{P(RT = \text{TRUE} \mid R = \text{TRUE})\,P(R = \text{TRUE})}{P\,(RT = \text{TRUE})}.$$

Ймовірність події $P(RT = \text{TRUE})$ можна обчислити, використовуючи усі можливі значення R і D з таблиць 1 і 2.

$$P(RT = \text{TRUE}) = P(D = \text{TRUE})P(R = \text{TRUE})P(RT \mid D = \text{TRUE}, R = \text{TRUE}) +$$

$$+ P(D = \text{FALSE})P(R = \text{TRUE})\,P(RT \mid D = \text{FALSE}, R = \text{TRUE}) +$$

$$+ P(D = \text{TRUE})\,P(R = \text{FALSE})P(RT \mid D = \text{TRUE}, R = \text{FALSE}) +$$

$$+ P(D = \text{FALSE})P(R = \text{FALSE})P(RT \mid D = \text{FALSE}, R = \text{FALSE}) =$$

$$= 0{,}05 \times 0{,}3 \times 0{,}9 + 0{,}95 \times 0{,}3 \times 0{,}5 + 0{,}05 \times 0{,}7 \times 0{,}2 + 0{,}95 \times 0{,}7 \times 0{,}1 = 0{,}2295.$$

Обчислимо ймовірність того, що товар повернено тому, що він бракований.

$$P(RT = \text{TRUE} \mid R = \text{TRUE}) = P(RT = \text{TRUE} \mid R = \text{TRUE}, D = \text{TRUE}) \times$$

$$\times P(D = \text{TRUE}) + P(RT = \text{TRUE} \mid R = \text{TRUE}, D = \text{FALSE})P(D = \text{FALSE}) =$$

$$= 0{,}9 \times 0{,}05 + 0{,}5 \times 0{,}95 = 0{,}52.$$

Тепер можемо обчислити ймовірність повернення товару через брак (рис. 5).

$$P(R = \text{TRUE} \mid RT = \text{TRUE}) = \frac{P(RT = \text{TRUE} \mid R = \text{TRUE})P(R = \text{TRUE})}{P(RT = \text{TRUE})} =$$

$$= \frac{0{,}52 \times 0{,}3}{0{,}2295} = 0{,}679739.$$

Отже, знання про те, що товар було повернуто, спричиняє перегляд усіх ймовірностей появи цієї події. Зміни відбулися в збільшенні ступеня упевненості, що повернення сталося через брак товару: від початкового значення 0,3 до 0,68.

Розглянемо можливість використання нашої методики для задачі кредитування фізичних осіб, яка останнім часом стає актуальною для України. Зараз комерційні банки починають усвідомлювати можливості розширення кредитування фізичних осіб, а для цього необхідно чітко передбачити всю необхідну інформацію про клієнтів. Обсяги кредитування вже почали збільшуватись, але почали збільшуватись і втрати банків від неповернення кредиту. За останні два роки прямі збитки банків від ведення кредитної діяльності становили біля 1,1 мільярда гривень [2].

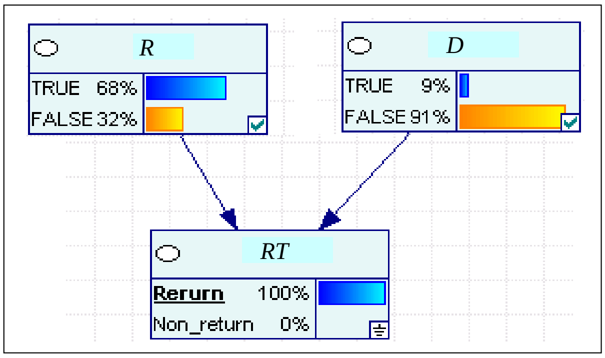

Рис. 5. Визначення ймовірності повернення товару через брак

Нехай іноземний комерційний банк B працює в Україні і видає кредити на купівлю побутової техніки, мобільних телефонів і т.п. та збирає статистику щодо повернення кредитів. За рік банком видано 1600 кредитів. Кожний клієнт описується 18 характеристиками (рис. 6). Задача формулюється наступним чином: 1) за наявними даними побудувати мережу, яка покаже зв'язок між характеристиками клієнта і вершиною — подією повернення кредиту; 2) за наявними навчальними даними визначити ймовірність повернення кредиту новим клієнтом.

Структуру відповідної МБ у вигляді гістограм наведено на рис. 7. З мережі видно, що освіта впливає на посаду клієнта, стать — на рівень його доходу, бо й досі зберігається тенденція, коли чоловіки працюють на керівних посадах та отримують більші доходи у порівнянні з жінками. Вік впливає на стаж роботи, а той — на строк проживання, область реєстрації клієнта — на область проживання, що вказана в паспорті. Існує також зв'язок між такими змінними, як вік, кількість дітей, сімейний стан та стать. Строк надання кредиту явно впливає на кредит та у подальшому на ймовірність його повернення так само, як і платіж, який необхідно сплачувати кожного місяця.

Судячи із структури побудованої мережі, безпосередній вплив на ймовірність повернення кредиту мають такі характеристики, як вік та сума кредиту. Це зрозуміло, бо люди віком до 18 та після 60 років взагалі не можуть отримати кредит, тому що законодавством не передбачена видача кредиту неповнолітнім, а банки не зацікавлені у видачі кредитів пенсіонерам. Побудована мережа дає ймовірність повернення кредиту лише на рівні 0,721659, близькому до граничного, тому діяльність банку з видачі кредитів важко вважати успішною.

Розглянемо приклад, коли до банку прийшов новий клієнт, який хоче отримати кредит. Нехай він — молодий чоловік віком 25 років з вищою освітою, бере кредит розміром 1000 гривень на один рік, тобто щомісяця сплачує 125 гривень, при цьому його дохід після усіх витрат за місяць становить 1000 гривень. Клієнт мешкає і зареєстрований у м. Києві 25 років, має власну квартиру, неодружений та не має дітей, працює на комерційному підприємстві зі штатом 15 осіб на посаді менеджера менше одного року (це його перша робота).

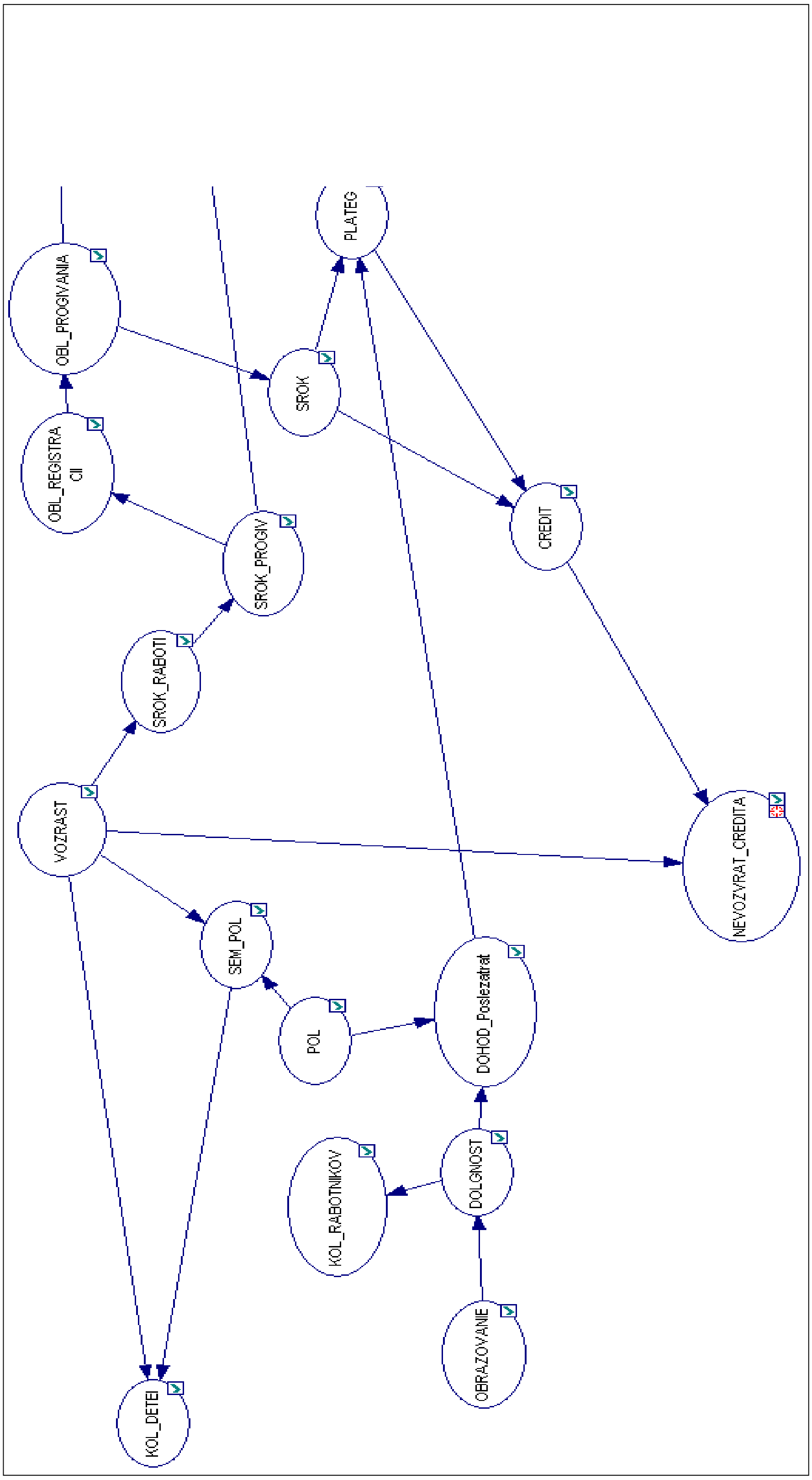

Рис 6. Структура мережі, побудованої за допомогою GeNie, для аналізу кредитів

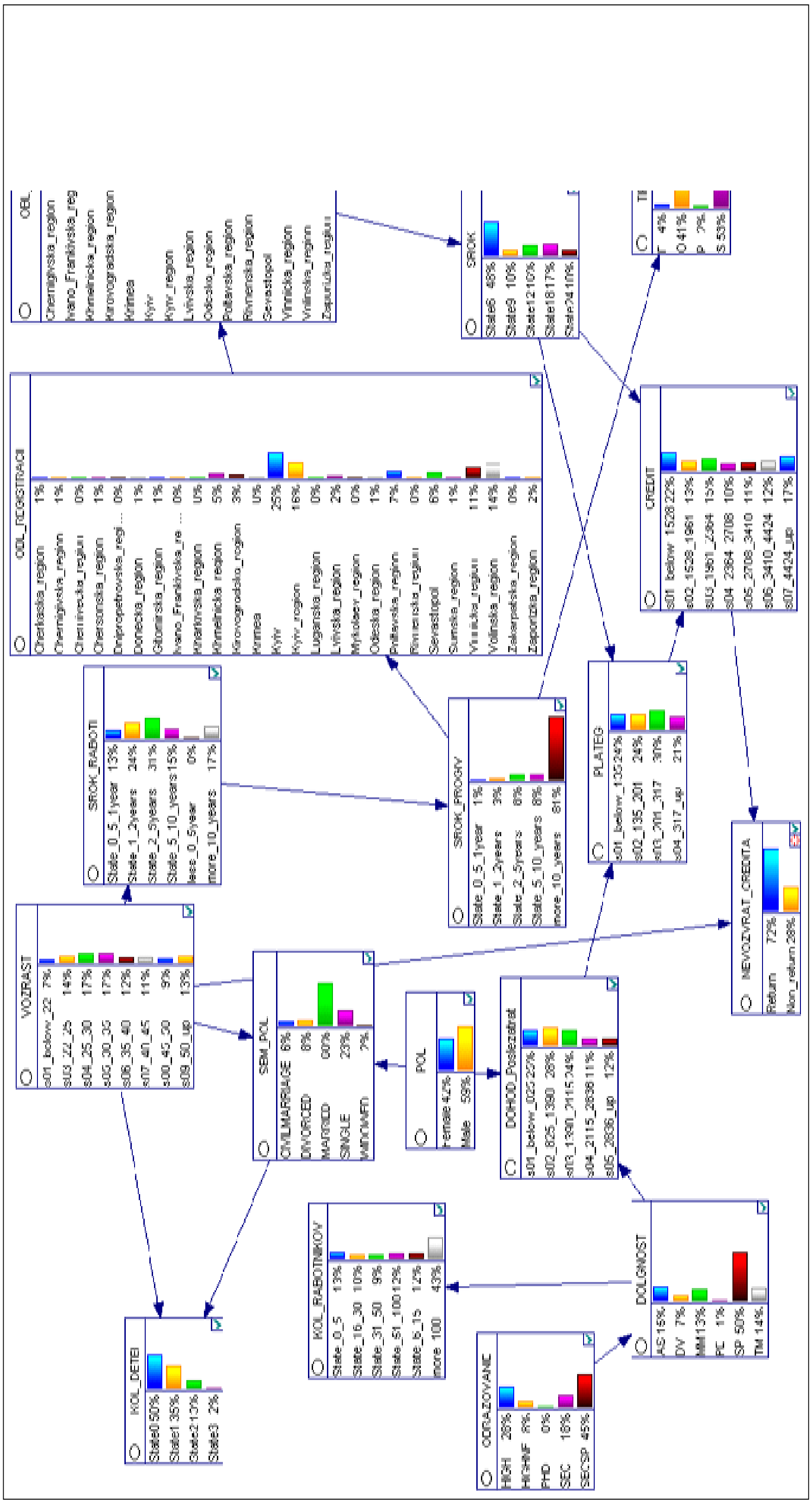

Рис. 7. Структура мережі, побудованої за допомогою GeNIe

Задавши всі ці дані клієнту в мережі, отримаємо, що ймовірність повернення кредиту цим клієнтом є доволі високою — 0,907407. За існуючими правилами банку цей клієнт може отримати кредит.

Аналіз мережі в даному випадку показує, що, можливо, доцільно внести зміни у внутрішню політику банку щодо видачі кредитів, змінити перелік відомостей, які надає клієнт про себе, започаткувати ведення кредитних історій.

ВИСНОВКИ

Проведений аналіз показав, що використання МБ стає дедалі частішим при побудові СППР, які мають працювати в умовах невизначеності. Розширення сфери застосування МБ сьогодні вимагає створення теоретичних основ та напрацювання алгоритмів автоматичної побудови МБ для конкретних випадків, накопичення даних для навчання і зменшення часу на одержання рішень.

Доцільно застосовувати апарат МБ для вирішення фінансових та економічних задач, яким притаманні випадковість, неповнота знань, складність предметної області.

ЛІТЕРАТУРА

1. *Бидюк П.И., Терентьев А.Н.* Построение и методы обучения байесовских сетей// Таврический вестник информатики и математики. — 2004. — № 2. — С. 139–153.
2. *Краснов С.О.* Неформальний аналіз кредитоспроможності індивідуальних позичальників комерційних банків // Наук. записки. — 2006. — Вип. 15. — 4 с.
3. *Сорокин А.В.* Сети Байеса // Сб. «Интеллектуальные технологии и системы». — М.: Изд-во МГУП, 1999. — Вып. 2. — 304 с.
4. *Хабаров С.П.* Экспертные системы. Конспект лекций. — 2001. // http://firm.trade.spb.ru/serp.
5. *Aronsky D., Haug P.J.* Diagnosing community-acquired pneumonia with a Bayesian network // Proc. AMIA Symp. — 1998. — P. 632–636.
6. *Dagum P., Luby M.* Appximating probabilistic inference in Bayesian belief networks is NP-hard // Artificial Intelligence. — 1993. — **45.** — P. 141–153.
7. *Lauritzen S.L., Spiegelhalter D.J.* Local computations with probabilities on graphical structures and their application to expert systems // Journal of the Royal Statistical Society. — Series B. — 1988. — **50**, № 2. — P.157–224.
8. *Murphy K.* A Brief Introduction to Graphical Models and Bayesian Networks. — 1998. — 19 p. // http://www.ai.mit.edu/~murphyk/Bayes/bnintro.html.
9. *Pearl J.* A constrained propagation approach to probabilistic reasoning // Kanal L.M., Lemmer J. (Eds.). — Uncertainty in Artificial Intelligence. — Amsterdam: North-Holland, 1986. — P. 357–370.
10. *Pearl J.* Probabilistic Reasoning in intelligent systems: networks of plausible inference. — San Mateo, CA (USA): Morgan Kauffmann Publishers, Inc., 1988. — 550 p.

Надійшла 22.06.2007