УДК 62-50

Згуровський М.З., Бідюк П.І., Терентьєв О.М.

Метод адаптування ймовірнісної байєсівської моделі до статистичних даних

Вступ. Популярним сучасним методом інтелектуального аналізу даних довільної природи є мережі Байєса (БМ). Байєсівська мережа – це ймовірнісна математична модель процесу (об'єкта), яка формується з двох елементів: (1) графічна структура та (2) ймовірнісна специфікація мережі у вигляді таблиць умовних ймовірностей. МБ можна розглядати як модель представлення причинних ймовірнісних залежностей (взаємозв'язків) між змінними (вершинами графа) досліджуваного процесу. Зв'язок $A \to B$ називають причинним, якщо подія A є причиною виникнення B, тобто якщо існує явний або прихований механізм впливу значень змінної A на значення змінної B. МБ називають причинною (каузальною) тоді, коли всі її зв'язки є причинними. БМ надають можливість будувати моделі процесів довільної природи в умовах наявності таких невизначеностей: (1) неповна інформація щодо процесу; (2) неможливість охопити всі тонкощі функціонування процесу внаслідок його високої розмірності; (3) наявність невимірюваних випадкових впливів та прихованих змінних процесу; (4) неявно виражена взаємодія між якісними і кількісними змінними і т. ін.

Формально, Байєсова мережа – це трійка $\mathbf{N} = <\mathbf{V}, \mathbf{G}, \mathbf{J}>$, першою компонентою якої є множина змінних $\mathbf{V}$; другою – спрямований ациклічний граф $\mathbf{G}$, вузли якого відповідають випадковим змінним модельованого процесу; J – спільний розподіл ймовірностей змінних $\mathbf{V} = \{X_1, X_2, ..., X_n\}$. При цьому стосовно множини змінних виконується марковська умова, тобто кожна змінна мережі не залежить від усіх інших змінних, за винятком батьківських попередників цієї змінної. Випадковість змінних процесу не виключає наявності в них детермінованої складової. Прозора гнучка графічна структура такої моделі і можливості формування ймовірнісного висновку за

допомогою різних методів на основі наявної та нової (додаткової) інформації зумовили можливості її різноманітного застосування в системах розпізнавання, класифікації, прогнозування, підтримки прийняття рішень.

Таким чином, графічна структура БМ задається орієнтованим ациклічним графом **G** , вузлами якого є змінні досліджуваного процесу, а дуги від батьківських вузлів до вузлів-нащадків представляють ймовірнісні взаємозв'язки між цими вузлами [1, 2]. Вузол X_j називають батьківським, або безпосереднім предком вузла, якщо існує дуга від X_j до X_i, що трактується так: „ X_j безпосередньо спричиняє появу X_i ". Надалі поняття «вузол» і «змінна» будемо вживати як синоніми.

Ймовірнісна специфікація задається розподілом умовних ймовірностей для значень кожного вузла X_i мережі відносно можливих ініціалізацій множини Π_i батьківських вузлів (безпосередніх попередників). Ініціалізація означає задавання початкових значень умовних ймовірностей вузлам на основі апріорної інформації. Для кореневих вузлів, тобто вузлів без вхідних дуг, визначається апріорний розподіл ймовірностей. Таким чином, невизначеність впливу батьків на нащадка описується за допомогою ймовірнісного взаємозв'язку. Однією з найпривабливіших можливостей даного математичного інструменту є чітка графічна інтерпретація результатів моделювання, яка інтуїтивно зрозуміла особі, що приймає рішення (ОПР). Графічна інтерпретація – це візуалізація причинно-наслідкових взаємозв'язків між елементами процесу, що моделюється за допомогою БМ.

Навчання МБ означає: (1) побудову структури мережі і (2) обчислення таблиць умовних ймовірностей для кожного вузла. Навчену мережу можна використати для формування точного або наближеного ймовірнісного висновку. Теоретичні аспекти навчання БМ викладені, зокрема, в [3, 4, 5]. Відомий алгоритм К2 навчання байєсівських мереж має доказову теоретичну базу і запроваджує вимоги, що забезпечують збільшення його швидкодії [6]. Однак даний алгоритм відноситься до алгоритмів типу «greedy search», тобто

його обчислювальні витрати можуть зростати поліноміально, оскільки в загальному випадку задача навчання БМ відноситься до NP-повних задач. А це означає, що застосування алгоритму К2 в системах реального часу, які вимагають перерахунку результатів після отримання нових даних, є можливим у разі достатнього запасу часу між моментами надходження послідовних пакетів даних.

Таким чином, існує об'єктивна необхідність адаптування результатів моделювання за допомогою К2 до нових даних. Метою такого підходу є зменшення обчислювальних витрат, оскільки алгоритм К2 обробляє весь пакет даних цілісно, що при неперервному збільшенні числа спостережень в базі даних призводить до експоненціального зниження продуктивності. Функціонал якості, покладений в основу К2, має властивість локальності обчислення, тобто його можна розкласти на складові, кожна з яких залежить від конфігурації вхідних дуг лише одного вузла мережі. Дану властивість буде використано для розробки алгоритму адаптування, який цілісно оброблятиме пакет нових даних.

Серед альтернативних підходів до задачі адаптування БМ найбільш фундаментальними є напрями, що грунтуються на використанні принципу опису мінімальної довжини (ОМД) [7, 8, 9], на використанні функціоналів якості, що дозволяють порівняти моделі розподілів із різних за об'ємами баз даних [10], а також на концепції часткових і повних теорій [11].

Для того щоб сформулювати постановку задачі введемо такі позначення: $Z = \{X_1,...,X_n\}$ – множина вузлів БМ, яка визначається числом змінних досліджуваного процесу в базі даних; $E = \{(X_i, X_j) | X_i, X_j \in Z\}$ – множина дуг мережі; X_i – вузол БМ, що відповідає спостереженням однієї змінної з бази даних; $n = |Z|$ – число вузлів БМ; r_i – число значень, які може приймати вузол X_i; v_{ik} – k-е значення змінної X_i; Π_i – множина вузлів-предків вузла X_i; ϕ_i – множина можливих ініціалізацій Π_i; $q_i = |\phi_i|$ – число можливих ініціалізацій Π_i; ϕ_{ij} – j-а ініціалізація множини вузлів-предків Π_i вузла X_i; B_S – структура БМ; B_P – ймовірнісна специфікація БМ, тобто частина опису моделі, що

представляє ймовірнісні характеристики БМ; $\theta_{ijk} = p(X_i = v_{ik} \mid \phi_{ij}, B_P)$ при цьому сума ймовірностей $\sum_k \theta_{ijk} = 1$; $f(\theta_{ij1},\ldots,\theta_{ijr_i})$ – щільність розподілу ймовірностей для вузла X_i та ініціалізації ϕ_{ij}; D_0 – вихідна база даних спостережень; S_0 – структура БМ, отримана внаслідок попередньої пакетної обробки бази D_0; D_1 – база даних нових спостережень, не використаних при побудові S_0; S_1 – структура БМ, отримана після адаптації S_0 до нових даних D_1.

Постановка задачі. Необхідно розробити алгоритм адаптування вихідної (первісної) байєсівської мережі $G = \langle Z, E \rangle$ із структурою S_0, побудованою за вихідною базою спостережень D_0, до нових спостережень D_1. Тобто ставиться задача формування оновленої структури мережі в зв'язку із зміною ситуації на об'єкті $S_1 \Leftrightarrow D_1$. При цьому експериментальні (статистичні) дані можуть мати довільний розподіл ймовірностей, а процеси, які описуються цими даними, мають, як правило, нестаціонарний характер, тобто математичне сподівання $M[X_i] \neq const$ і дисперсія $M\{X_i - M[X_i]\}^2 \neq const$.

Для розв'язку задачі скористаємось наступними вихідними даними:

- вихідні дані спостережень, що відповідають вимогам припущенням алгоритму К2;
- байєсівська нейронна мережа, побудована в результаті попередньої пакетної обробки бази спостережень за алгоритмом К2;
- нові спостереження, які не використовувалися при попередньому формуванні байєсівської мережі.

Загальні припущення та положення алгоритму адаптування. Оскільки метод адаптації байєсівської мережі передбачає використання функціоналу якості, який покладено в основу алгоритму К2 навчання БМ, то необхідно розглянути початкові припущення цього алгоритму [6].

1. Для побудови моделі використовуються дискретні змінні із попередньо сформованої бази даних.
2. Спостереження, накопичені в базі даних, є незалежними одне від одного. Структура ймовірнісної моделі, за якою могли бути отримані ці спостереження, є незмінною впродовж всього часу спостережень.

   У разі, коли між фактичними змінними існує деякий взаємозв'язок, можна збільшити число спостережень, тобто додати в базу даних додаткові змінні, які описують ці взаємозв'язки.
3. Кожний сеанс спостережень є повним, тобто містить інформацію про всі спостережувані змінні. Дана вимога може бути дещо пом'якшена [6], але цей випадок необхідно необхідно розглядати окремо.
4. Функції щільності розподілу ймовірностей $f(\theta_{ij1},\ldots,\theta_{ijr_i})$ і $f(\theta_{i_0j_01},\ldots,\theta_{i_0j_0r_i})$ є незалежними для всіх $1\le i, i_0\le n$, $1\le j\le q$, $1\le j_0\le q_0$, $ij\ne i_0j_0$. Це припущення можна представити у вигляді:

$$f(B_P \mid B_S)=\prod_{i=1}^{n}\prod_{j=1}^{q_i} f(\theta_{ij1},\ldots,\theta_{ijr_i}).$$

5. Функції $f(\theta_{ij1},\ldots,\theta_{ijr_i})$ представляють рівномірний розподіл, тобто $f(\theta_{ij1},\ldots,\theta_{ijr_i})=C_{ij}$, де $C_{ij}=const$. Цей факт інтерпретується так: ми не надаємо переваги у виборі значень умовних ймовірностей $\theta_{ij1},\ldots,\theta_{ijr_i}$ якійсь одній множині значень у порівнянні з іншою множиною.

Визначимо додаткові аспекти розв'язання задачі адаптування структури мережі, зумовлені особливостями алгоритму К2. Ці аспекти визначають концептуальний підхід до розв'язання задачі адаптування, а також деякі практичні питання реалізації відповідного алгоритму.

6. Задача адаптування байєсівської мережі розглядатиметься з використанням байєсівського підходу по аналогії із задачею навчання БМ для алгоритму К2.
7. За функціонал якості використовується функція, покладена в основу К2. Властивість локальності даної функції буде враховуватися при побудові алгоритму адаптування.

8. Вимогою до вхідної інформації є збереження задавання послідовності вузлів БМ, що визначає причинно-наслідкові зв'язки між вузлами.

Функціонал якості алгоритму К2, який використовується при побудові структури мережі, має вигляд:

$$p(D \mid B_P) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}!,$$

де $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$; N_{ijk} – число спостережень із бази даних D, в яких $\{X_i = v_{ik}; \pi_i = \phi_{ij}\}$. Розглянемо випадок отримання множини нових спостережень розміром $|D_1| = M$. Для кожного вузла X_i беремо вибірку спостережень із D_1, в якій $\{X_i = v_{ik}; \Pi_i = \phi_{ij}\}$. Число елементів такої вибірки $\dim\{X_i = v_{ik}; \Pi_i = \phi_{ij}\} = m_{ijk}$. Оскільки спостереження повні, то $\sum_{j,k} m_{ijk} = M$.

Розглянемо випадок, коли дані містять лише M однакових спостережень. Нехай окреме спостереження ініціює вузол X_i значенням v_{ik_0} , а множина вузлів-предків Π_i задається набором ϕ_{ij_0} . Тепер

$$p(D_1, D_0 \mid S_0) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! =$$

$$= \prod_{i=1}^{n} \frac{(r_i - 1)!}{(N_{ij_0} + r_i - 1)!} \prod_{\substack{j=1 \\ j \neq j_0}}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{\substack{k=1 \\ k \neq k_0}}^{r_i} N_{ijk}! \cdot N_{ij_0k_0}! =$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \times$$

$$\times \prod_{i=1}^{n} \left[\frac{1}{N_{ij_0} + r_i} \cdots \frac{1}{N_{ij_0} + r_i + M - 1} \right] \times \left[(N_{ij_0k_0} + 1) \cdots (N_{ij_0k_0} + M) \right] =$$

$$= p(D_0 \mid S_0) \cdot \prod_{i=1}^{n} \frac{\prod_{u=1}^{M} \left(N_{ij_0k_0} + u \right)}{\prod_{u=1}^{M} \left(N_{ij_0} + r_i - 1 + u \right)}.$$

Але з іншого боку, $p(D_1, D_0 \mid S_0) = p(D_0 \mid S_0) \cdot p(D_1 \mid D_0, S_0)$, тому отримаємо:

$$p(D_1 \mid D_0, S_0) = \prod_{i=1}^{n} \frac{\prod_{u=1}^{M} \left(N_{ij_0k_0} + u\right)}{\prod_{u=1}^{M} \left(N_{ij_0} + r_i - 1 + u\right)}.$$

Аналогічні викладки для випадку довільної множини спостережень в сукупності D_1 дають такий результат:

$$p(D_1 \mid D_0, S_0) = \prod_{i=1}^{n} \frac{\prod_{s=1}^{R_i} \prod_{t=1}^{Q_i} \prod_{u=1}^{m_{its}} \left(N_{its} + u\right)}{\prod_{t=1}^{Q_i} \prod_{u=1}^{M_{it}} \left(N_{it} + r_i - 1 + u\right)},$$

де R_i – число значень змінної X_i, які містилися в пакеті D_1; Q_i – число значень змінної X_i, що з'явилися в пакеті даних ініціалізації множини вузлів-предків Π_i; $M_{it} = \sum_k m_{itk}$. Як зазначалось вище, це значення можна інтерпретувати як вклад нових даних в процес навчання БМ. Відповідно, для коефіцієнта K_{delete} отримуємо формулу:

$$K_{delete}(i) = \frac{\prod_{s=1}^{R_i} \prod_{t=1}^{Q_i} \prod_{u=1}^{m_{its}} \left(N_{its} + u\right)}{\prod_{t=1}^{Q_i} \prod_{u=1}^{M_{it}} \left(N_{it} + r_i - 1 + u\right)}.$$

Цей коефіцієнт необхідний для ретроспективного аналізу впливу нових спостережень на процес навчання БМ. Таким чином, використовуючи локальність функціоналу якості алгоритму К2, одержуємо індикатори, необхідні для адаптування тієї або іншої дуги. При цьому структура адаптаційного алгоритму К2 залишається незмінною.

Адаптаційний варіант алгоритму К2. Адаптування побудованої мережі виконується в такій послідовності:

1. Реалізація процедури корегування структурної частини моделі:
   1.1. Процедура видалення дуг, які не узгоджуються з множиною даних.
   1.2. Процедура додавання нових дуг.
2. Процедура корегування ймовірнісної частини моделі.

Оскільки на початковому етапі навчання БМ ймовірнісну складову моделі представляють таблиці умовних розподілів ймовірностей, отримані безпосередньо на підставі частотного аналізу появи значень змінних в спостереженнях, тому відразу визначимо зміни у процедурі корегування ймовірнісної частини моделі.

З метою полегшення процедури корегування ймовірнісної частини моделі корисно зберігати не таблиці розподілу умовних ймовірностей, а значення N_{ijk}. Це дозволить швидше оновлювати дані щодо розподілу умовних ймовірностей, а самі значення умовних ймовірностей можна обчислити за формулою Діріхле:

$$p(X_i = v_{ik} \mid \Pi_i = \phi_{ij}) = \frac{N_{ijk} + 1}{N_{ij} + r_i}.$$

При корегуванні структури БМ порядок обходу вузлів визначаємо за вкладом кожного вузла в значення: $p(D_1 \mid D_0, S_0) = \prod_{i=1}^{n} \frac{\prod_{s=1}^{R_i} \prod_{t=1}^{Q_i} \prod_{u=1}^{m_{its}} (N_{its} + u)}{\prod_{t=1}^{Q_i} \prod_{u=1}^{M_{it}} (N_{it} + r_i - 1 + u)}$.

Аналіз інформаційної важливості дуг встановлюється так. На етапі перевірки дуг на можливість видалення для кожного вузла обчислюється значення $K_{delete}(S_0)$ для поточної конфігурації множини вузлів-предків, а також значення $K_{delete}(S_{-1}^m)$ для конфігурацій, які представляють собою результат видалення однієї з M $(1 \le m \le M)$ вхідних дуг з поточного вузла. Якщо виконується умова $K_{delete}(S_{-1}^m) \le K_{delete}(S_0)$, то m-а дуга залишається в структурі мережі, оскільки видалення цієї дуги призводить до зменшення значення локального функціоналу якості (тобто для поточного вузла). Інакше, дуга заноситься в список дуг, що підлягають подальшій перевірці на необхідність видалення. Список може бути відсортований за збільшенням значення $K_{delete}(S_{-1}^m)$. Список (множина) дуг аналізується послідовно. Подальша перевірка полягає в обчисленні значення локального функціоналу якості при вихідній конфігурації і конфігураціях, які ми отримуємо при видаленні однієї з дуг, що залишилися у переліку.

Тактика вилучення і додавання дуг застосована у інкрементному варіанті адаптаційного алгоритму, наведеному нижче. Оскільки результатом реалізації байєсівського підходу є вибір стратегії адаптування у вигляді

$$P(S_1 \mid D_1, D_0, S_0) = \arg\max_S \frac{P(S \mid D_0) P(D_1 \mid S, D_0)}{P(D_1 \mid S_0, D_0)},$$

то процедури вилучення і додавання дуг будемо здійснювати так, як зазначено нижче.

Якщо врахувати вид розв'язку оптимізаційної задачі адаптування БМ, то тактика вилучення дуг повинна приводити до зменшення першої складової чисельника $P(S \mid D_0)$, оскільки вона досягає максимуму при $S = S_0$ в результаті формування початкової структури БМ. Таким чином, для отримання позитивного ефекту від адаптування необхідно компенсувати втрати від вилучення дуги ефектом від додавання нової дуги. Оскільки вихідною умовою алгоритму К2 є наявність впорядкованої послідовності вузлів, то пошук дуги-претендента на додавання здійснюється саме в такому порядку. Оцінка внеску дуги виконується шляхом обчислення значення локального функціоналу якості. Відповідно, претендент на додавання повинен визначати конфігурацію вхідних дуг, що має найбільше значення локального функціоналу якості.

Після кожного видалення виконується спроба додати можливу дугу з набору предків претендентів, в який включається також предок видаленої дуги. Якщо в результаті такої спроби виявляється, що видалена дуга максимально збільшує значення локального функціоналу якості у порівнянні з іншими претендентами, то ця дуга повертається назад в структуру і виконується перехід до оцінювання значення функціоналу після видалення наступної дуги.

Якщо після видалення дуги, першою дугою, вибраною для додавання, є невидалена дуга, то ітерація додавання виконується до тих пір, поки не настане одна з таких ситуацій:

1) всі претенденти додані, включаючи видалену дугу, яка була додана останньою;

2) на черговій ітерації вибрана дуга, що була видалена раніше, повертається назад в структуру;
3) на черговій ітерації не можна підвищити значення локального функціонала якості шляхом додавання нової дуги із множини предків-претендентів; якщо цей результат наступає при аналізі останньої дуги з претендентів на видалення, то для даного вузла немає потреби виконувати кроки 1, 2.

Для вузлів, що не мають дуг-претендентів на видалення, виконується тільки перевірка необхідності додавання нової дуги. Відповідно до алгоритму К2 процедура додавання дуг виконується для кожного вузла згідно з порядком його місцезнаходження у вхідній послідовності вузлів. Процедура продовжується до тих пір, поки жодна з дуг, що залишилися, не буде збільшувати локальне значення функціонала якості.

Моделювання і загальний план тестування. Вихідна база даних створена за допомогою СУБД Access. Оскільки послідовність вузлів, що утворюють вхідні дані для алгоритму К2, суттєво впливає на якість моделі, то заздалегідь будується раціональна конфігурація предків для кожного вузла. За критерій якості використано максимум функціоналу якості алгоритму К2. На підставі отриманих локальних конфігурацій вибрано відповідну послідовність вузлів, що дає можливість одержати кращі локальні конфігурації мережі в ході реалізації алгоритму К2. Від цього кроку можна було б відмовитися, оскільки алгоритм дозволяє визначити кращу конфігурацію мережі в межах заданої послідовності вузлів. Тобто всі етапи тестування були однаково обмежені первісно обраною послідовністю.

На етапі перевірки алгоритму адаптування початкова база даних розбивалася на 2 частини: перша («старі» спостереження) використовувалася для попереднього навчання мережі, а друга («нові» спостереження) – для додаткового навчання.

Загальний план тестування методу складається з таких кроків:

1. Побудова моделі за алгоритмом К2 на основі бази «нових» спостережень з метою визначення залежностей, що містяться в нових даних.
2. Побудова моделі за алгоритмом К2 на основі бази «старих» спостережень з метою визначення залежностей, що містяться в даних, використовуваних для початкового навчання мережі.
3. Створення моделі за алгоритмом К2 за базою «старих» спостережень з її подальшим адаптуванням до «нових» спостережень.
4. Побудова моделі за алгоритмом К2 на основі об'єднаної бази «старих» і «нових» спостережень з метою порівняння результуючої мережі з результатом кроку 3.
5. Попередня ініціалізація мережі, яка є неадекватною базі «старих» спостережень, з її подальшим адаптуванням до «нових» спостережень з метою порівняння результуючої мережі з результатом виконання кроку 3.

Використані дані. При моделюванні використано дані у вигляді деперсоніфікованої історії авторизованих запитів за допомогою платіжних банківських карток окремих клієнтів. За цими картками зафіксовано шахрайські операції. Мета моделювання полягала у побудові байєсівської мережі, що відображає характеристики поведінки клієнтів при використанні платіжних карток. Надалі ця модель може використовуватися для розв'язання задачі класифікації нових авторизацій, здійснюваних за допомогою даної картки з метою виявлення шахрайських операцій.

Оригінальні дані попередньо оброблялись з метою спрощення процесу моделювання. Всі дані, окрім грошових еквівалентів, трактувалися як дискретні змінні. Для обробки грошових еквівалентів запроваджено інтервали фактичних сум з метою дискретизації відповідних змінних. Після обробки структура даних, використаних при моделюванні, складалась з полів, наведених у табл. 1.

Таблиця 1

Перелік і опис полів бази даних, використаної в процесі моделювання

Назва поля бази даних	Кількість можливих значень змінної, відповідної даному полю	Логічний зміст значення поля
KEY	Лічильник	Фактично визначає унікальний ключ кожного запису бази.

FRAUD	2 можливі значення	Визначає, чи являється дана операція шахрайської.
DAYSBEFO	8 можливих значень	Визначає кількість днів між даною операцією і попередньої.
ACCNT_AMT	6 можливих інтервалів сум	Визначає суму поточної операції.
SPENT_AM	6 можливих інтервалів сум	Визначає загальну суму всіх операцій за поточний день.
TRAN_CCY	5 можливих значень	Визначає код валюти, в якій здійснена поточна операція.
MCCCLASS	10 можливих значень	Визначає код виду діяльності торгівельної точки, в якій здійснена поточна операція (тобто те, який товар/послугу надає даний підприємець).
CITY	10 можливих значень	Визначає місто, в якому здійснена дана операція.
COUNTRY	4 можливі значення	Визначає країну, в якій здійснена дана операція.

Всього використано 434 повних спостережень.

Результати тестування методу. На попередньому етапі вибрано наступну послідовність вузлів-змінних, яка узгоджується з вимогами даного алгоритму:

0. COUNTRY; 1. TRAN_CCY; 2. ACCNT_AMT; 3. FRAUD; 4. CITY; 5. MCCCLASS; 6. DAYSBEFO; 7. SPENT_AM.

Тестування методу виконано для різних співвідношень між об'ємами баз «старих» і «нових» спостережень при групуванні початкової бази на інтервали: 300-43 і 173-170. Результати виявилися практично однаковими, тому далі буде описано результати тестування для першого розподілу вибірки даних. Результати побудови структури мережі в ході першого і другого етапу тестування наведені у табл. 2.

Таблиця 2

Результати етапів 1 і 2 тесту: кращі структури мереж, значення локальних та інтегральних функціоналів якості

Етап 1: Побудова моделі за алгоритмом К2 з використанням бази «нових» спостережень з метою визначення залежностей, що містяться в нових даних, за якими адаптується мережа			Етап 2: Побудова моделі за алгоритмом К2 з використанням бази «старих» спостережень з метою визначення залежностей, що містяться в даних, за якими виконується попереднє навчання мережі		
Назва вузла	Номери предків даного вузла	Значення локального функціонала якості для даної конфігурації предків даного вузла	Назва вузла	Номери предків даного вузла	Значення локального функціонала якості для даної конфігурації предків даного вузла

0. COUNTRY [] -5,8147402271 4 значення	0. COUNTRY [] -71,7851105248 4 значення
1. TRAN_CCY [0] -5,9116502401 5 значень	1. TRAN_CCY [0] -22,6209092243 5 значень
2. ACCNT_AMT [] -17,498550055 6 значень	2. ACCNT_AMT [] -140,339677935 6 значень
3. FRAUD [0] -10,006491539 2 значення	3. FRAUD [] -2,47856649559 2 значення
4. CITY [3,0] -23,280635627 10 значень	4. CITY [1,0] -101,149752148 10 значень
5. MCCCLASS [3,0] -32,053370070 10 значень	5. MCCCLASS [1,2,0] -175,887625370 10 значень
6. DAYSBEFO [] -26,710288145 8 значень	6. DAYSBEFO [] -149,699813640 8 значень
7. SPENT_AM [6] -19,740909612 6 значень	7. SPENT_AM [6] -143,708728638 6 значень
----------	----------
Log10(P)=-141,016635517887	log10(P)=-807,670183977427

Головна відмінність між мережами, побудованими на етапах 1 і 2, зумовлена тим, що шахрайські операції з картками, які визначають вузол FRAUD, входили в базу «нових» спостережень і відсутні в базі «старих» спостережень. З урахуванням даної поправки встановлено, що обидві структури достатньо точно описують логіку причинно-наслідкових зв'язків між вузлами і не суперечать експертним висновкам.

Структури мереж, одержаних на етапах 3 і 4, співпадають, що свідчить про адекватність функціонування алгоритму адаптації. Інформаційний опис мереж наведено у табл. 3. Єдиною відмінністю даних звітів є послідовність розташування вузлів-предків для вузлів CITY і MCCCLASS. Це пояснюється тим, що розташування вузлів-предків в даних послідовностях визначається ходом алгоритму адаптації (для етапу 3) і ходом алгоритму К2 (для етапу 4). Оскільки на етапі 3 адаптація починається з моменту побудови мережі у відповідності зі «старою» базою спостережень, то початковою мережею для адаптації є структура, що співпадає з результатом етапу 2. Таким чином, результати етапу 2 стають відправною точкою для адаптування. З урахуванням вищезгаданої різниці між «старими» і «новими» даними, стає зрозуміло, що саме відмінність логіки адаптації (етап 3) від логіки пакетного навчання (етап 4)

визначає орієнтацію результатів етапу 3 на результат етапу 2. При цьому дотримуються характеристики відповідності між результатами етапів 3 і 4. На рис. 1 наведена структура мережі, яка відповідає результату виконання етапів 3 і 4.

Таблиця №3

Результати етапів 3 і 4: кращі структури мережі, значення локальних і інтегральних функціоналів якості

Етап 3: Побудова моделі алгоритмом К2 по базі «старих» спостережень з подальшою її адаптацією до «нових» спостережень			Етап 4: Побудова моделі алгоритмом К2 по об'єднаній базі «старих» і «нових» спостережень		
Назва вузла	Номери предків даного вузла	Значення локального функціонала якості для певної конфігурації предків даного вузла	Назва вузла	Номери предків даного вузла	Значення локального функціонала якості для певної конфігурації предків даного вузла
0. COUNTRY 4 значення	[]	-78,1985152776	0. COUNTRY 4 значення	[]	-78,1985152776
1. TRAN_CCY 5 значень	[0]	-23,6660822012	1. TRAN_CCY 5 значень	[0]	-23,6660822012
2. ACCNT_AMT 6 значень	[]	-155,732839411	2. ACCNT_AMT 6 значень	[]	-155,732839411
3. FRAUD 2 значення	[]	-19,7484793156	3. FRAUD 2 значення	[]	-19,7484793156
4. CITY 10 значень	[1,0,3]	-118,470260543	4. CITY 10 значень	[1,3,0]	-118,470260543
5. MCCCLASS 10 значень	[1,0,2,3]	-202,469443226	5. MCCCLASS 10 значень	[1,2,3,0]	-202,469443226
6. DAYSBEFO 8 значень	[]	-172,299888064	6. DAYSBEFO 8 значень	[]	-172,299888064
7. SPENT_AM 6 значень	[6]	-155,708893197	7. SPENT_AM 6 значень	[6]	-155,708893197
----------			----------		
log10(P)=-926,294401238406			log10(P)=-926,294401238406		

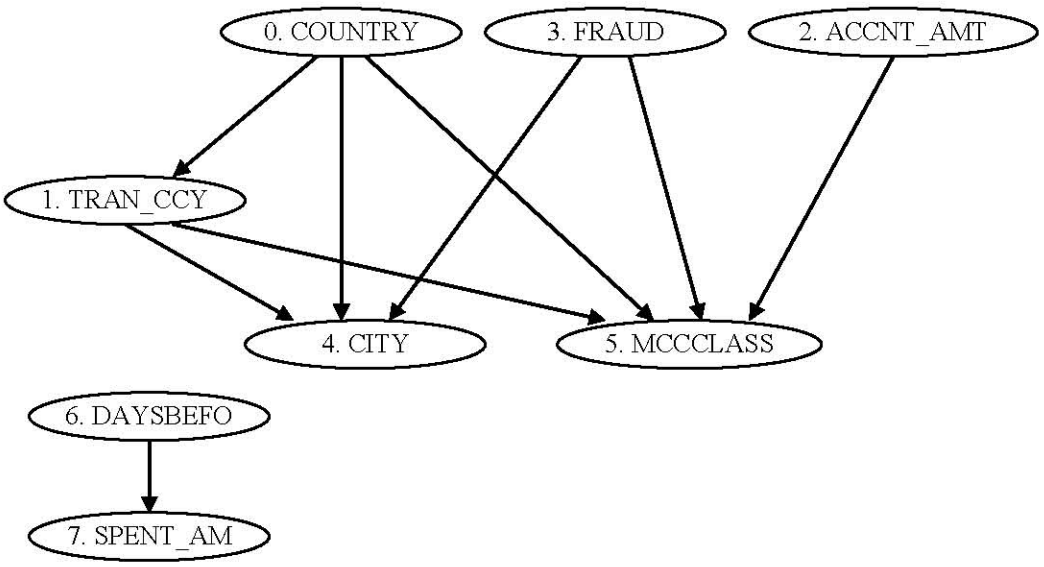

Рис. 1. Структура мережі Байєса

На етапі 5 зроблено спробу аналізу ефективності адаптування в умовах недостовірної інформації щодо структури мережі. На першому кроці мережу ініціалізовано вручну деякою структурою, що містила декілька дуг, неадекватних (або слабо адекватних) базі «старих» спостережень (для порівняння див. табл. 2). Результати адаптування наведено в табл. 4.

Таблиця 4

Результати етапу 5: початкова ініціалізація мережі недостовірною інформацією, результат адаптації, значення локальних і інтегральних функціоналів якості

Етап 5, крок ініціалізації мережі недостовірною структурою			Етап 5, крок адаптації ініціалізованої мережі до «нових» даних		
Назва вузла	Номери предків даного вузла	Значення локального функціонала якості для даної конфігурації предків даного вузла	Назва вузла	Номери предків даного вузла	Значення локального функціонала якості для даної конфігурації предків даного вузла
0. COUNTRY 4 значення	[]		0. COUNTRY 78,1985152776 4 значення	[]	-
1. TRAN_CCY 5 значень	[0]		1. TRAN_CCY 23,6660822012 5 значень	[0]	-
2. ACCNT_AMT 6 значень	[1,0]		2. ACCNT_AMT 161,556768690 6 значень	[1,0]	-
3. FRAUD 2 значення	[1,0,2]		3. FRAUD 20,7471725966 2 значення	[1,0]	-
4. CITY 10 значень	[1,0]		4. CITY	[1,0,3]	-
5. MCCCLASS 10 значень	[1,2,0]				

6. DAYSBEFO [] 8 значень 7. SPENT_AM [6] 6 значень	118,470260543 10 значень 5. MCCCLASS [1,0,2,3] - 202,469443226 10 значень 6. DAYSBEFO [] - 172,299888064 8 значень 7. SPENT_AM [6] - 155,708893197 6 значень -------------------------- log10(P)=-933,117023799001

Як видно з отриманих результатів адаптування, серед свідомо помилкових дуг-взаємозв'язків була видалена тільки одна – від ACCNT_AMT до FRAUD. Це свідчить про те, що попереднє навчання мережі за базою «старих» спостережень є необхідним етапом для отримання адекватних результатів адаптації, оскільки дана вимога покладена в основу теоретичної розробки алгоритму адаптації на основі К2.

Висновки. Мережі Байєса – гнучкий та зручний інструмент моделювання складних, процесів, які характеризуються дискретними та неперервними змінними і функціонують в умовах наявності невизначеностей різних типів. Запропоновано метод навчання та адаптування мережі Байєса до нової інформації, який грунтується на базових ідеях алгоритму К2 навчання байєсівських мереж і забезпечує додаткові можливості щодо адаптування структури попередньо сформованої БМ до нових даних. Необхідно зазначити, що попередньо сформована мережа не обов'язково повинна мати високу ступінь адекватності до даних, оскільки в подальшому реалізується процедура повторного навчання, яка спрямована на рекурсивне уточнення її структури за новими даними з довільним розподілом та статистичними характеристиками. При цьому нові дані формуються в пакети, що забезпечують достатньо ефективне навчання мережі.

Застосування декомпозиції функціоналу якості з подальшим ретроспективним аналізом якості мережі дозволяє отримати критерії, які необхідні для процедури адаптації моделі до нових даних. Отримані результати комп'ютерного імітаційного моделювання підтвердили можливість використання розробленого пакетного алгоритму для адаптації попередньо сформованої байєсівської мережі, побудованої за базою початкових спостережень, до нових спостережень.

В подальших дослідженнях необхідно встановити вплив розмірів (об'ємів) пакетів даних на якість навчання мережі та синтезувати статистичну міру відмінності нових пакетів даних від попередніх з метою встановлення необхідності повторного навчання. Також доцільно розглянути задачу скорочення обчислювальних витрат на навчання.

Література

1. Cooper W., Maron M. Foundations of probabilistic and utility-theoretic indexing // Journal of the ACM, 1978, Vol. 25, pp. 67-80.
2. Murphy K. A Brief Introduction to Graphical Models and Bayesian Networks. http://www.cs.berkeley.edu/~murphyk/Papers/intel.ps.gz. – 19 p.
3. Heckerman D. A tutorial on learning Bayesian networks. Technical report MSN-TR-95-06, Microsoft Research, Advanced Technology Division, 1995. – 52 p.
4. Бидюк П.И., Терентьев А.Н., Гасанов А.С. Построение и методы обучения байесовских сетей // Кибернетика и системный анализ, 2005, № 4, с. 133-147.
5. Терентьєв О.М. Застосування SPI-методу для формування ймовірносного висновку в мережах Байєса / О.М. Терентьєв, П.І. Бидюк / Математичне та програмне забезпечення інтелектуальних систем: зб. наук. праць за матеріалами V міжнар. наук.-прак. конф., 14-16 лист. 2007 р., Дніпропетровськ. – Дніпропетровськ: ДНУ, 2007. – С. 189-190.
6. Jordan M.J. Graphical models / Technical report, 2000. – 20 p.
7. Zgurovsky M.Z., Bidyuk P.I., Terentyev O.M. Methods of constructing Bayesian networks based on scoring functions // Cybernetics and Systems Analysis, 2008, Vol. 44, No. 2, pp. 219-224.
8. Терентьев А.Н. Эвристический метод построения байесовских сетей / А.Н. Терентьев, П.И. Бидюк / Интеллектуальные системы принятия решений и прикладные аспекты информационных технологий: сб. науч. трудов по материалам II междунар. конф., 15-18 мая 2006 г., Евпатория. – Херсон: ХМИ, 2006.– Т. 1. – С. 401–403..
9. Lam W., Bacchus F.. Using new data to refine Bayesian network. *Proceeding of the Tenth Conference on Uncertainty in Artificial Intelligence*, 1994, pp. 383-390.
10. Friedman N., Goldszmidt M.. Sequential update of Bayesian network structure // *Proceeding of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, 1997, p. 273-280.
11. Buntine W. Theory refinement on Bayesian networks. *Proceeding of the Seventh Conference on Uncertainty in Artificial Intelligence*, p.p. 52-60, 1997.

Zgurovsky M.Z., Bidyuk P.I., Terentyev O.M. The Method for adapting probabilistic Bayesian model to statistical data // A method for learning and adapting Bayesian network structure to new data is proposed that is based on application of Bayesian approach and K2 algorithm. The method provides additional possibilities for adapting the preliminary formed network structure to new data and saving computing time.

Згуровский М.З., Бидюк П.И., Терентьев А.Т. Метод адаптации вероятностной байесовской модели к статистическим данным // Предложен метод обучения и адаптации структуры сети Байеса к новой информации, который основывается на использовании байесовского подхода и базовых предположений алгоритма К2. Метод обеспечивает дополнительные возможности для адаптирования структуры предварительно сформированной БС к новым данным и снижение вычислительных затрат на обучение БС.

Згуровський М.З., Бідюк П.І., Терентьєв О.М. Метод адаптування ймовірнісної байєсівської моделі до статистичних даних // Запропоновано метод навчання і адаптування структури мережі Байєса до нової інформації, який грунтується на використанні байєсівського підходу та базових припущень алгоритму К2. Метод забезпечує додаткові можливості щодо адаптування структури попередньо сформованої БМ до нових даних і зменщення обчислювальних витрат на навчання мережі.