ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ, СИСТЕМНИЙ АНАЛІЗ ТА КЕРУВАННЯ

DOI: 10.20535/kpi-sn.2020.3.209877
УДК 519.766.4, 519.25

Л.Б. Левенчук, П.І. Бідюк*
КПІ ім. Ігоря Сікорського, Київ, Україна
*corresponding author: pbidyuke_00@ukr.net

БАЙЄСІВСЬКИЙ АНАЛІЗ ДАНИХ У МОДЕЛЮВАННІ ТА ПРОГНОЗУВАННІ НЕЛІНІЙНИХ НЕСТАЦІОНАРНИХ ПРОЦЕСІВ

Проблематика. Нелінійні нестаціонарні процеси, що виникають у різних сферах діяльності людини, пов'язані з великою кількістю невизначеностей, нечіткістю, неповнотою та неточністю даних. Для прогнозування таких процесів необхідно коректно опрацьовувати дані такого типу, тому актуальною є задача розробки і застосування нових методів, які дають можливість здійснювати належну обробку вхідних даних з метою моделювання і прогнозування досліджуваних процесів.
Мета дослідження. Виконати короткий огляд методів байєсівського аналізу даних, розробити методику ідентифікації та врахування можливих невизначеностей у моделюванні й прогнозуванні, а також запропонувати комплексну імовірнісно-статистичну модель для прогнозування нелінійних нестаціонарних процесів.
Методика реалізації. Комплексно застосовано: байєсівський метод обробки даних, оптимальний фільтр для попередньої обробки даних та їх підготовки до побудови моделей, лінійну та нелінійну регресійні моделі для формального опису і прогнозування умовної дисперсії та ймовірнісну модель у формі байєсівської мережі для прогнозування нелінійного нестаціонарного процесу.
Результати дослідження. Запропонований метод моделювання апробовано на задачі оцінювання прогнозів фінансового процесу на фондовому ринку. Використані статистичні дані описують еволюцію цін на акції для відомої компанії. В результаті виконання обчислювальних експериментів було встановлено, що якість короткострокових прогнозів волатильності й самого нелінійного нестаціонарного процесу значно поліпшується завдяки оптимальній фільтрації даних і раціональній структурі моделі. Застосування побудованої комплексної моделі з використанням байєсівської мережі надало можливість удосконалити ймовірнісне оцінювання прогнозів при здійсненні торговельних операцій на фондовому ринку.
Висновки. Оцінювання прогнозів нелінійних нестаціонарних процесів є актуальною задачею, що може розв'язуватися різними методами. Високоефективним виявився запропонований ймовірнісно-статистичний метод для оцінювання ймовірнісних прогнозів при здійсненні торговельних операцій на фондовому ринку акцій, а тому в подальшому перспективним буде розширення та удосконалення його застосування.

Ключові слова: байєсівський аналіз даних; нелінійні нестаціонарні процеси; фільтр Калмана; мережа Байєса; регресійна модель; комбінування прогнозів.

Вступ

Застосування ймовірнісно-статистичних методів необхідно розглядати як важливу трудомістку частину процесу прийняття рішень, яка дає можливість приймати обґрунтовані тактичні та стратегічні рішення. Ці методи ґрунтуються на глибокому аналізі наявної інформації, а також на знаннях, досвіді та інтуїції фахівців (експертні оцінки). Ймовірнісний аналіз процесів, подій і даних різних типів передбачає два підходи [1–3]:

– частотний, який ґрунтується на ймовірнісному аналізі за класичним підходом – дані накопичуються у процесі виконання експериментів (або збору статистики) і обробляються так званими частотними методами теорії ймовірностей;

– байєсівський, в основу якого покладається той чи інший варіант теореми Байєса; цей підхід не виключає використання класичних методів, а вхідна інформація для аналізу може бути подана у вигляді статистичних (експериментальних) даних, експертних оцінок, окремих фактів тощо.

Коректне використання ймовірнісно-статистичних методів надає значні переваги стосовно прийняття коректних рішень практично в усіх напрямах людської діяльності: в конкуренції за підвищення якості та продаж нової продукції, дає можливість отримати високоякісні оцінки прогнозів, обґрунтувати фінансові, макроекономічні рішення, рішення стосовно раціонального ведення домашнього господарства, а також розв'язати багато інших задач. Послідовність реалізації етапів розв'язання задачі мо-

делювання і подальшого застосування побудованої моделі при використанні ймовірнісно-статистичних методів залишається такою ж, як і при використанні інших методів, наприклад регресійного аналізу [4–6].

Весь процес моделювання і подальше застосування побудованої моделі складаються з таких етапів:

– вивчення теоретичних основ можливих методів моделювання, придатних для розв'язання задач конкретного типу;

– вибір методу (методів) моделювання;

– збір даних і побудова моделей-кандидатів, що належать до вибраного класу моделей; вибір кращої моделі з множини побудованих кандидатів за статистичними критеріями адекватності;

– обчислення оцінок прогнозів за допомогою побудованої моделі;

– використання прогнозів для автоматичного (чи іншого типу) керування або для підтримки прийняття управлінських, особистих та інших типів рішень.

Постановка задачі

Робота присвячена розв'язанню таких задач: (1) огляд деяких методів байєсівського аналізу даних; (2) аналіз невизначеностей, що зустрічаються у процесі обробки статистичних/експериментальних даних з метою побудови математичних моделей досліджуваних процесів, оцінювання прогнозів та синтезу систем керування; (3) ілюстрація практичного застосування вибраного методу байєсівського аналізу даних.

Методологія байєсівського програмування

Сьогодні байєсівська методологія моделювання та формування ймовірнісного висновку (остаточного результату) на основі моделі отримала назву *байєсівського програмування*. Методологія означає, що для аналізу даних використовується множина методів, які забезпечують розв'язання таких задач [7, 8]:

– побудова ймовірнісно-статистичних моделей різних типів (оцінювання структури та параметрів) із використанням статистичних даних і експертних оцінок;

– обчислення остаточних результатів на основі створеної моделі згідно з постановкою задачі: оцінок прогнозів, керуючих впливів, оцінок змінних і параметрів на виході фільтрів, розпізнавання образів, знаходження рішень стосовно управління досліджуваними процесами і об'єктами тощо;

– аналіз коректності отриманих результатів за відповідними множинами статистичних критеріїв якості;

– практичне застосування отриманих результатів та їх корегування (повторне оцінювання) залежно від остаточного результату.

До методології байєсівського програмування відносять методи, описані нижче. Рекурсивне *байєсівське оцінювання*: *фільтрація*, *прогнозування*, *згладжування* змінних; основне рівняння оцінювання має такий вигляд:

$$P(S(k) \mid O(0)\ldots O(k)) = P(O(k) \mid S(k)) \times$$
$$\times \sum_{S(k-1)} [(P(S(k) \mid S(k-1))(P(S(k-1) \mid O(0)\ldots O(k-1)],$$

де $S(0),\ldots, S(k)$ – часовий ряд змінних стану; $O(0),\ldots, O(k)$ – часовий ряд спостережень; $P(S(k) \mid S(k-1))$ – модель системи або модель переходів; $P(O(k) \mid S(k))$ – модель спостережень, що показує, яким буде спостереження в момент k, якщо система перебуває в стані $S(k)$.

За цією моделлю ми можемо знайти $P(S(k+l) \mid O(0)\ldots O(k))$ – тобто яким буде розподіл ймовірностей станів у момент $k+l$, якщо маємо спостереження $O(0),\ldots, O(k)$. Якщо $l = 0$, то реалізується процедура фільтрації; якщо $l > 0$, то реалізується процедура прогнозування; а при $l < 0$ відбувається згладжування – відновлення минулого стану на основі спостережень, зроблених до або після моменту згладжування.

Приховані марковські моделі (ПММ) – це модифікація байєсівського фільтра, в якому припускається, що дані дискретні; моделі переходів станів і спостережень задаються матрицями ймовірностей або *таблицями умовних ймовірностей*. Якщо спостережувані змінні неперервні, то такі моделі називають *напівнеперервними* ПММ.

Оптимальні рекурсивні *фільтри Калмана* (КФ); змінні неперервні або дискретні; моделі переходів станів і спостережень задаються з використанням гауссівських процесів – зовнішніх випадкових збурень і похибок (шумів) вимірів. У випадку нелінійних моделей використовують розклад у ряд Тейлора, що уможливлює лінійно локальні моделі. Для одночасного оцінювання

станів досліджуваних процесів та їх параметрів використовують розширений фільтр Калмана (РКФ).

Гранулярні (particle) *фільтри* (ГФ); розподіл ймовірностей станів описують такою моделлю: $P(S(k-1) \mid O(0),...,O(k-1))$ (тут $S(\cdot)$ – матриця станів; $O(\cdot)$ – матриця спостережень), що апроксимується множиною гранул (particles), вагові коефіцієнти яких пропорційні ймовірностям їх появи. Для оновлення ймовірностей станів використовується рекурсивна процедура.

Статичні *байєсівські мережі* (static Bayesian networks (BN)) – це ймовірнісно-статистичні моделі для опису ймовірнісної та статистичної інформації в умовах наявності невизначеності. На змінні мережі практично не накладаються обмеження, і немає спеціальної семантики для їх опису – тобто існує певна свобода вибору змінних для побудови мережі. Змінними мережі можуть бути дискретні та неперервні змінні, а також експертні оцінки, зведені до необхідної числової форми.

Байєсівська мережа (БМ) подається у формі спрямованого ациклічного графа, вершинами якого є змінні досліджуваного процесу, а дуги вказують на існуючі умовні залежності між змінними. Формально мережу можна описати трійкою: $BN = \{V, G, T\}$, де V – змінні (дані) для побудови мережі (база даних); G – спрямований ациклічний граф; T – таблиця умовних ймовірностей для вершин графа (змінних моделі).

Параметрами такої моделі є умовні ймовірності в таблицях умовних ймовірностей. Для батьківських (незалежних) змінних це таблиці безумовних ймовірностей. Неперервні змінні подаються відповідними розподілами. Якщо БМ містить дискретні та неперервні змінні, її називають гібридною. Неперервні змінні в таких випадках, як правило, дискретизують, що дає можливість суттєво спростити обчислювальні операції.

Послідовність побудови моделі у формі БМ можна подати у вигляді таких кроків:

(1) поглиблений аналіз досліджуваного процесу (об'єкта) з метою встановлення особливостей його функціонування та виявлення батьківських і дочірніх змінних;

(2) виявлення існуючих моделей процесу й аналіз можливості їх подальшого використання в ІСППР;

(3) встановлення існуючих зв'язків між змінними процесу за допомогою спеціальних тестів та експертного оцінювання;

(4) скорочення розмірності задачі моделювання;

(5) масштабування і дискретизація змінних;

(6) визначення семантичних обмежень для моделі;

(7) оцінювання структур моделей-кандидатів із використанням оптимізаційних процедур, тобто пошук альтернативних моделей у формі БМ;

(8) аналіз якості та вибір кращої з моделей-кандидатів;

(9) застосування вибраної моделі для розв'язання поставленої задачі;

(10) формування ймовірнісних висновків стосовно вибраних змінних за побудованою моделлю (моделями), аналіз якості отриманого результату.

Результатом побудови і використання БМ є ймовірнісний висновок у формі $P(X^i \mid Known)$, де $Known$ – підмножина інших змінних мережі, ймовірності станів яких відомі на момент обчислення ймовірнісного висновку. Загалом ймовірнісний висновок у БМ полягає в поширенні ймовірностей і параметрів гауссівських законів розподілу по всій мережі залежно від отриманих *свідчень* (додаткової інформації про стани мережі). В основі процесу формування ймовірнісного висновку покладено досить складні математичні алгоритми.

Динамічні байєсівські мережі (dynamic Bayesian networks (DBN)) створюються для того, щоб враховувати динаміку процесів (їх зміни в часі), а також можливі стохастичні впливи на їх перебіг. Фактично, динамічна байєсівська мережа (ДБМ) – це розширення звичайних (статичних) мереж. Спочатку будується звичайна БМ для наявних змінних, структура якої передбачається інваріантною стосовно часу, тобто залишається сталою. Така структура повторюється для кожного наступного моменту часу з надходженням нових спостережень. Таким чином досягається відтворення динаміки (змін у часі) досліджуваних процесів.

Частину графа, що відповідає конкретному моменту часу t_k або просто k, називають часовим перерізом. Якщо приймається гіпотеза стосовно того, що стан поточного часового перерізу залежить тільки від попереднього, то таке припущення називають *марковським припущен-*

ням першого порядку. Якщо структура всіх часових перерізів однакова, то таку ДБМ називають *стаціонарною*. У такому випадку модель, що відповідає одному часовому перерізу, називають *локальною* та *інваріантною стосовно часу* або *гомогенною*.

Моделі *марковської локалізації* (Markov localization (ML) models) – це моделі типу байєсівських фільтрів, які додатково включають керуючі змінні $\mathbf{u}(0),\ldots,\mathbf{u}(k-1)$. Іноді їх називають ще прихованими марковськими моделями за входом-виходом.

У такій моделі ймовірності станів уточнюються за допомогою керуючих змінних таким чином: $P(S(k)\mid \mathbf{u}(k-1),S(k-1))$. Таку модель називають ще моделлю відтворення дії (action model). Форми таких моделей можуть бути різними. Загалом це матричні моделі; а якщо вони подібні до тих, які використовуються в гранулярних фільтрах, то їх називають моделями Монте-Карло з марковською локалізацією (Monte Carlo Markov Localization (MCML)).

Побудована модель призначена для того, щоб давати відповідь на запитання стосовно ймовірності поточного стану досліджуваного об'єкта

$$P(S(k)\mid \mathbf{u}(0),..,\mathbf{u}(k-1),O(0),\ldots,O(k))$$

на основі попередніх керуючих дій та спостережень за динамікою об'єкта. Термін "локалізація" пов'язаний із застосуванням у робототехніці, тобто розглядається задача локалізації (місцезнаходження) робота в навколишньому середовищі.

Основне рівняння моделі подібне до основного рівняння фільтрації і має вигляд

$$\begin{aligned} P(S(k)\mid \mathbf{u}(0),\ldots,\mathbf{u}(k-1),(0)\ldots O(k)) &= \\ = P(\mathbf{u}(k-1)P(O(k)\mid S(k))\times & \\ \times \sum_{S(k-1)} [(P(S(k)\mid \mathbf{u}(k-1),S(k-1))(P(S(k-1) & \\ \mid \mathbf{u}(0),.....,\mathbf{u}(k-2),\ O(0)\ldots O(k-1)]. & \end{aligned}$$

Крім названих вище, широко використовуються такі методи:

– *байєсівські карти* – це узагальнення моделей марковської локалізації, які також виникли у сфері керування роботами;

– байєсівський метод обробки даних і *прийняття рішень* на основі *ієрархічних моделей*;

– байєсівська регресія та узагальнені лінійні моделі.

Схема байєсівського підходу до моделювання та прийняття рішень

Байєсівський підхід до побудови моделей і формування ймовірнісного висновку передбачає виконання етапів, зображених на рис. 1 [8, 9]. У байєсівському аналізі даних передбачається, що інформація надходить із двох джерел: апріорна інформація від дослідника стосовно досліджуваної задачі та нові статистичні (експериментальні) дані, отримані в результаті виконання експериментів.

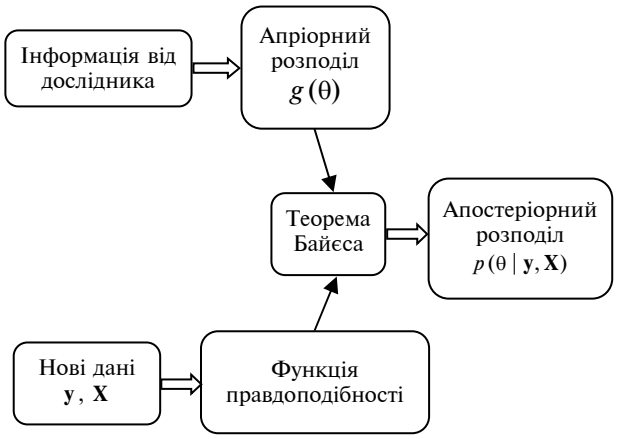

Рис. 1. Інформаційні потоки в системі моделювання на основі теореми Байєса

Апріорна інформація від дослідника відображає результати попередніх досліджень; теоретичні основи досліджуваних явищ, процесів та об'єктів; додаткові неформалізовані дані, отримані з різних джерел. Загалом – це додаткова інформація стосовно досліджуваних процесів (об'єктів), яка не повинна ґрунтуватись на експериментальних даних, що стосуються конкретної задачі.

У центрі байєсівської методології аналізу даних знаходиться теорема Байєса (ТБ). Запишемо її з урахуванням позначень, що широко використовуються в економетричному аналізі даних:

$$p(\theta\mid \mathbf{y},\mathbf{X}) = \frac{g(\theta)\,p(\mathbf{y}\mid \theta,\mathbf{X})}{p(\mathbf{y})} \propto g(\theta)\,p(\mathbf{y}\mid \theta,\mathbf{X}), \quad (1)$$

де $\mathbf{y}$ – вектор вимірів основної (залежної) змінної досліджуваного процесу; $\mathbf{X}$ – матриця вимірів незалежних змінних (регресорів), які визначають поведінку основної змінної; θ – вектор випадкових параметрів, які (разом із $\mathbf{X}$) ви-

значають функцію щільності розподілу ймовірностей змінної y (у класичній регресійній моделі θ – вектор параметрів регресійної моделі); $p(\cdot)$ – функція щільності розподілу ймовірностей; $p(\theta)$ – апріорна щільність розподілу ймовірностей для значень випадкового параметра θ, яка ґрунтується на апріорних знаннях дослідника до використання даних $(\mathbf{y}, \mathbf{X})$; $p(\mathbf{y} \mid \theta, \mathbf{X})$ – умовна щільність даних $\mathbf{y}$ за конкретних значень θ і $\mathbf{X}$, іншими словами, це функція правдоподібності для даних $\mathbf{y}$; $p(\mathbf{y})$ – *маргінальна правдоподібність* для $\mathbf{y}$ (оскільки в процесі обчислень $p(\mathbf{y})$ вилучено вплив незалежних змінних і параметрів); символ $\propto$ означає "наближення".

Щільність $p(\theta \mid \mathbf{y}, \mathbf{X})$ – апостеріорний розподіл θ, який ґрунтується на оновленні апріорного розподілу завдяки даним $(\mathbf{y}, \mathbf{X})$. Тобто ТБ показує, як можна об'єднати інформацію, що надходить із двох джерел (апріорний розподіл і дані) з метою уточнення апріорного розподілу. Фактично, ТБ не тільки дає можливість об'єднувати інформацію з двох джерел: її суть у тому, що будь-яке інше комбінування даних буде порушувати логічну (і математичну) сутність правил оперування з розподілами ймовірностей.

У виразі (1) перший варіант теореми записано як рівність, оскільки у ньому наявний знаменник $p(\mathbf{y})$ (безумовна щільність для $\mathbf{y}$), який відіграє роль *нормуючої константи* і забезпечує те, що апостеріорна умовна щільність для θ є *належною* та інтегрується до 1 по області визначення параметрів. Другий варіант ТБ у виразі (1) подано з точністю до константи пропорційності, тобто без нормування, що часто робиться для спрощення подання виразів. При розв'язанні практичних задач спочатку обчислюється чисельник, а потім, за необхідності, нормуюча константа. У багатьох прикладах існує необхідність в обчисленні відношень результатів, а тому нормуюча константа скорочується.

Кількість потоків інформації, що подаються на ТБ з метою обчислення апостеріорного розподілу параметра θ, може бути більшою одиниці. Тобто на вході може бути вектор змінних. Відносний вплив двох і більше джерел інформації на остаточний результат залежить від точності їх представлення. Наприклад, чим меншою буде дисперсія апріорного розподілу, тим більшу роль (з точки зору точності остаточного результату) він відіграватиме у формуванні апостеріорного розподілу.

Боротьба з невизначеностями у байєсівському моделюванні

У загальному випадку *невизначеність* – це *фактор негативного впливу* на процес попередньої обробки даних, побудови математичної моделі, оцінювання прогнозу та генерування рішення на його основі, який призводить до погіршення якості проміжних та остаточних результатів.

Поява невизначеності зумовлена недостатністю або спотворенням даних на будь-якому етапі згаданого процесу. Це може бути коротка вибірка даних, якої недостатньо для побудови адекватної моделі; спотворення вимірів шумовими складовими (збурення станів і похибки вимірів); некоректність встановлення типу розподілу даних і, як наслідок, неправильний вибір методу оцінювання параметрів моделі; помилки дослідника при виборі методу обчислення оцінок прогнозів або альтернативних рішень на їх основі.

У загальному випадку для боротьби з невизначеностями використовують нечітку логіку, ймовірнісно-статистичне моделювання, оптимальну і цифрову фільтрацію даних, методи заповнення пропусків і обробки екстремальних значень, альтернативні методи оцінювання параметрів тощо. Деякі типи невизначеностей, причини її виникнення та методи подолання подані в табл. 1. При побудові математичних моделей на основі даних зустрічаються три основних типи невизначеностей: статистична, структурна і параметрична [5, 10, 11].

Статистична невизначеність зумовлена самими даними, тобто наявністю пропусків вимірів, короткою вибіркою, наявністю екстремальних значень, впливом зовнішніх випадкових збурень на досліджуваний процес і похибками вимірів. У деяких випадках виникає необхідність оцінювати невимірювані компоненти вектора стану досліджуваної системи (змінні), які мають значення для побудови адекватної математичної моделі. Для того щоб досягти прийнятних за точністю результатів оцінювання значень таких змінних, необхідно застосовувати альтернативні методи оцінювання, комплексування та комбінування вимірів.

Структурна невизначеність виникає при оцінюванні структури моделі на основі даних. Так, наприклад, оцінити порядок авторегресії

можна тільки наближено, оскільки значення автокореляційної функції, а також часткової автокореляційної функції – це випадкові величини. Це ж стосується оцінювання лагу (часу запізнення) по входу, типу нелінійності й типу нестаціонарності (які оцінюються за відповідними статистиками). На кожному етапі аналізу даних ми маємо справу з оцінками параметрів і змінних, які є випадковими величинами, що вносить похибки у проміжні та остаточні результати обчислень. Тому виникає необхідність ідентифікації таких невизначеностей і вибору (створення) методів мінімізації їх впливу на результати аналізу даних.

Параметрична невизначеність є наслідком наявності двох попередніх типів невизначеності. Вона зумовлена наближеними оцінками параметрів, що характеризують структуру моделі (порядок, значення запізнення по входу, оцінки параметрів розподілу), наявністю випадкових збурень і похибок вимірів. Досить часто неможливо встановити коректно тип розподілу даних унаслідок наявності коротких або спотворених вибірок – все це призводить до зміщення (зсуву) оцінок параметрів моделі від точних значень і підвищення дисперсії цих оцінок (відхилення від ефективності). Тому до вибору методу оцінювання параметрів на основі наявних даних

***Таблиця* 1.** Типи невизначеностей у моделюванні та методи їх подолання

Тип невизначеності	Причини виникнення	Методи подолання
Структурна невизначеність моделі	– неможливість встановлення всіх причинно-наслідкових зв'язків між змінними; – наближені значення елементів структури моделі	– експертні методи; – застосування статистичних тестів; – застосування теорії перевірки гіпотез; – байєсівські мережі, узагальнені лінійні моделі
Статистичні невизначеності в процесі побудови моделей	– похибки вимірів; – стохастичні зовнішні збурення; – мультиколінеарність; – екстремальні значення; – пропуски вимірів	– цифрові та оптимальні фільтри; байєсівські фільтри; – уточнення типів спільних розподілів; – метод головних компонент; – теорія екстремальних значень; – методи заповнення пропусків
Параметричні невизначеності	– некоректний вибір методу оцінювання; – короткі вибірки; – похибки вимірів; – випадкові зовнішні збурення	– забезпечення вибору альтернативних методів оцінювання параметрів; – метод Монте-Карло для марковських ланцюгів; – розмноження вибірок
Ймовірнісні невизначеності	– складні механізми виникнення причинно-наслідкових зв'язків; – відсутність детермінованості	– статичні та динамічні мережі Байєса; – марковські моделі; – ймовірнісні фільтри; – умовні багатовимірні розподіли
Невизначеність амплітудного типу	– наявність змінних, що не вимірюються	– методи обробки нечіткої інформації; – байєсівські мережі

завжди необхідно підходити обґрунтовано, враховуючи існуючі обмеження кожного методу.

Деякі методи подолання (мінімізації впливу) невизначеності при моделюванні на основі статистичних даних подані в табл. 1. У першу чергу це методи ймовірнісно-статистичного моделювання, нечітка логіка, оптимальна і цифрова фільтрація, методи обробки екстремальних значень та інші. З табл. 1 видно, що байєсівські методи моделювання можна застосовувати для боротьби практично з усіма типами невизначеностей. Найчастіше ймовірнісно-статистичне моделювання надає можливість успішно ідентифікувати, описати і врахувати невизначеності ймовірнісного, статистичного й амплітудного типів.

Невизначеності амплітудного типу зустрічаються у випадку наявності невимірюваних змінних або ж коли змінні не можна виміряти з необхідною точністю. У таких випадках для опису значень змінних та їх взаємодії можна застосовувати нечітку логіку або БМ. Який метод приведе до кращого остаточного результату, як правило, наперед невідомо, оскільки в кожного з них є свої переваги та недоліки. З практики застосування цих методів до мінімізації впливу невизначеності амплітудного типу відомо, що їх застосування дає можливість суттєво покращити проміжні та остаточні результати обчислювальних експериментів із використанням статистичних/експериментальних даних.

Приклад застосування байєсівського методу моделювання

Розглянемо побудову і застосування комбінованої моделі для короткострокового прогнозування нелінійного нестаціонарного фінансового процесу ціноутворення на біржі. Вибірка даних містила 320 значень процесу, 300 з яких було використано для навчання моделі, а 20 – для перевірки результату прогнозування. Один із варіантів структури використаної моделі подано на рис. 2. На етапі підготовки даних до побудови моделі передбачено заповнення пропусків альтернативними методами, а також оптимальну (фільтр Калмана) і цифрову фільтрацію вимірів [10, 11]. Зокрема, заповнення пропусків можливе такими методами: просте усереднення (за умови, що кількість пропусків не перевищує три), оптимізаційна процедура на основі ЕМ-алгоритму та обчислення прогнозних значень на основі лінійних регресійних моделей. Оскільки для застосування оптимального фільтра Калмана необхідно мати модель процесу в просторі станів, то за таку модель використано спрощений варіант: модель випадкового кроку, яка не потребує оцінювання параметрів:

$$x(k) = x(k-1) + w(k),$$

але враховує збурення стану $w(k)$. Рівняння вимірів, $z(k) = x(k) + v(k)$, враховує похибки вимірів $v(k)$. Незважаючи на простоту використаної моделі, оптимальна фільтрація дала можливість підвищити якість оцінок короткострокових прогнозів.

У випадку прогнозування самого процесу ціноутворення лінійна регресія використана для формального опису і прогнозування лінійної складової процесу, а нелінійна регресія і логістична функції забезпечують прогнозування нелінійної складової з її подальшим урахуванням у процедурі комбінування.

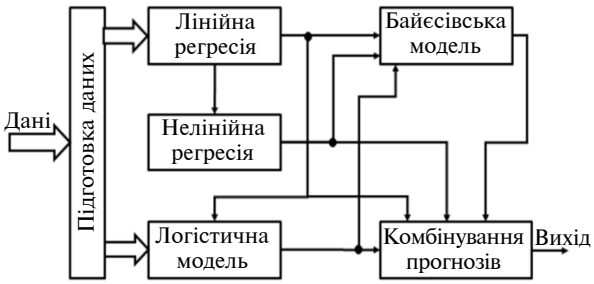

Рис. 2. Комбінована модель для прогнозування нелінійних процесів

Байєсівське програмування представлене у цьому випадку БМ. Вона призначена для обчислення ймовірнісної оцінки прогнозу, яка дає додаткову інформацію про динаміку процесу. Побудована модель дає можливість комбінувати оцінки прогнозів з метою подальшого підвищення якості прогнозу основної змінної. Результати однокрокового прогнозування вибраного процесу ціноутворення подані в табл. 2, а якість прогнозування волатильності цього процесу – в табл. 3.

Кращий результат однокрокового прогнозування (САПП = 5,38 на навчальній вибірці без фільтрації даних і САПП = 5,04 з фільтрацією) отримано за допомогою моделі, що складається з лінійної авторегресії п'ятого порядку і тренду такого ж порядку. В результаті комбінування оцінок прогнозів (три кращих методи) досягнуто значення САПП = 5,18 на навчальній вибірці та САПП = 5,19 на перевірочній вибірці.

***Таблиця* 2**. Прогнозування ціноутворення акцій Microsoft

Тип моделі	САПП без фільтрації	САПП із фільтрацією	Комбінований прогноз: АР(5) + t5, лог.+ лінійна, БМ
АР(5) (1)	8,47	7,95	–
АР(5) + t5 (1)	5,38	5,04	–
Логістична + лінійна регр. (1)	6,73	6,11	–
БМ (1)	7,49	7,32	5,18
АР(5) (2)	9,26	8,87	–
АР(5) + t5 (2)	5,76	5,43	–
Логістична + лінійна регр. (2)	7,15	6,64	–
БМ (2)	7,73	7,41	5,19

Примітка. САПП – середня абсолютна похибка в процентах; АР – авторегресія; БМ – байєсівська мережа.

***Таблиця* 3**. Прогнозування волатильності прибутку акцій Microsoft

Тип моделі	САПП без фільтрації	САПП із фільтрацією	Комбінований прогноз для ЕУАРУГ і МСВ
АРУГ (1)	9454,4	9188,7	–
УАРУГ (1)	45,960	36,270	–
ЕУАРУГ (1)	4,9400	3,7530	–
МСВ (1)	7,5000	4,9023	3,2475
АРУГ (2)	5123,5	2494,4	–
УАРУГ (2)	51,993	28,396	–
ЕУАРУГ (2)	5,93	4,0710	–
МСВ (2)	10,90	6,9590	3,4583

Примітка. САПП – середня абсолютна похибка в процентах; АРУГ – модель авторегресії з умовною гетероскедастичністю; УАРУГ – узагальнена АРУГ; ЕУАРУГ – експоненціальна УАРУГ; МСВ – модель стохастичної волатильності.

Статистичні параметри якості короткострокового прогнозування волатильності на навчальній (1) та перевірочній (2) вибірках подані в табл. 3. Досліджено модель авторегресії з умовною гетероскедастичністю (АРУГ), узагальнену АРУГ (УАРУГ), експоненціальну УАРУГ (ЕУАРУГ) і модель стохастичної волатильності (МСВ). Значення середньої абсолютної похибки у процентах (САПП), отримані на навчальній вибірці для моделей УАРУГ (1), ЕУАРУГ (1), МСВ (1), є дещо меншими, ніж для оцінок прогнозів, обчислених на перевірочній вибірці (УАРУГ (2), ЕУАРУГ (2) і МСВ (2)). Це був очікуваний результат. Найпростіша модель АРУГ продемонструвала велику неточність оцінок прогнозів на обох вибірках. Це пояснюється простотою її структури, яка не відображає фактичної взаємодії відповідних змінних.

Висновки

Поданий вище короткий огляд методів байєсівського програмування свідчить про наявність досить широкого спектра ймовірнісних методів для обробки статистичних/експериментальних даних і отримання результату у вигляді ймовірнісного висновку. Байєсівські методи дають можливість застосовувати оптимальну фільтрацію даних і будувати математичні моделі з використанням даних різних типів з їх подальшим використанням для прогнозування та підтримки прийняття рішень. Особливо популярні сьогодні БМ, які мають низку таких переваг: допускається висока розмірність моделей; одночасно можна використовувати дискретні та неперервні дані, а також експертні оцінки; БМ забезпечують оптимізаційний пошук фактичних взаємозв'язків між змінними; існує множина методів побудови оптимальної структури мережі та формування ймовірнісного висновку на її основі. БМ може бути побудована і використана автономно або у складі комбінованої моделі, яка включає, крім БМ, моделі інших типів. Поданий приклад прогнозування нелінійного нестаціонарного процесу ціноутворення за допомогою комбінованої моделі свідчить про можливість досягнення високої якості оцінок короткострокових прогнозів як самого процесу, так і його волатильності. Додаткового підвищення якості оцінок прогнозів можна досягти завдяки введенню в інформаційну систему обробки даних процедур оптимальної фільтрації та комбінування оцінок прогнозів, обчислених за аль-

тернативними методами. Це підтверджено виконаними обчислювальними експериментами.

У подальших дослідженнях доцільно розглянути такі задачі: застосувати ймовірнісну фільтрацію даних як частину інформаційної технології моделювання і прогнозування; використати розширену номенклатуру моделей для формального опису і прогнозування нелінійних нестаціонарних процесів, а також створити відповідну практичну методику моделювання і прогнозування процесів згаданого типу.

References

[1] P.D. Hoff, *A First Course in Bayesian Statistical Methods*. London, UK: Springer-Verlag, 2009, pp. 31–65.

[2] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. San Francisco: Morgan Kaufmann Publishers, 1988, pp. 77–141.

[3] S.J. Press, *Subjective and Objective Bayesian Statistics*. Hoboken: Wiley-Interscience, 2002, pp. 264–280.

[4] M.Z. Zgurovskii and P.I. Bidyuk, *Bayesian Networks in Decision Support Systems*. Kyiv, Ukraine: Edelweiss, 2015, pp. 12–23.

[5] S.O. Dovgiy *et al.*, *Decision Support Systems Based on Statistical and Probabilistic Methods*. Kyiv, Ukraine: Logos, 2014, pp. 94–107.

[6] P.I. Bidyuk *et al.*, *Mathematical Statistics*. Kyiv, Ukraine: Personal, 2018, pp. 138–243.

[7] B. Chen, "A Bayesian sampling approach to decision fusion using hierarchical models", *IEEE Trans. Signal Process*. vol. 50, no. 8, pp. 1809–1818, 2002. doi: 10.1109/TSP.2002.800419

[8] J.H. Dorfman, *Bayesian Economics Through Numerical Methods: A Guide to Econometrics and Decision Making with Prior Information*. New York: Springer-Verlag, 1997, pp. 88–96.

[9] W.R. Gilks *et al.*, *Markov Chain Monte Carlo in Practice*. Boca Raton: Chapman & Hall/CRC, 1995, pp. 45–59.

[10] A.C. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge, UK: University of Cambridge, 1990, pp. 100–167.

[11] R.S. Tsay, *Multivariate Time Series Analysis with R and Financial Applications*. Hoboken: Wiley-Interscience, 2014, pp. 279–319.

Л.Б. Левенчук, П.И. Бидюк

БАЙЕСОВСКИЕ МЕТОДЫ АНАЛИЗА ДАННЫХ В МОДЕЛИРОВАНИИ И ПРОГНОЗИРОВАНИИ НЕЛИНЕЙНЫХ НЕСТАЦИОНАРНЫХ ПРОЦЕССОВ

Проблематика. Нелинейные нестационарные процессы, которые возникают в различных сферах деятельности человека, связаны с большим количеством неопределенностей, нечеткостью, неполнотой и неточностью данных. Для прогнозирования таких процессов необходимо корректно обрабатывать данные такого типа, поэтому актуальной задачей является разработка и применение новых методов, которые дают возможность осуществлять надлежащую обработку данных с целью моделирования и прогнозирования исследуемых процессов.

Цель исследования. Выполнить краткий обзор методов байесовского анализа данных, разработать методику идентификации и учета возможных неопределенностей в моделировании и прогнозировании, а также предложить комплексную вероятностно-статистическую модель для прогнозирования нелинейных нестационарных процессов.

Методика реализации. Комплексно использованы: байесовский метод обработки данных, оптимальный фильтр для предварительной обработки данных и их подготовки к построению моделей; линейная и нелинейная регрессионные модели для формального описания и прогнозирования условной дисперсии и вероятностная модель в форме байесовской сети для прогнозирования нелинейного нестационарного процесса.

Результаты исследования. Предложенный метод моделирования апробирован на задаче оценивания прогнозов финансового процесса на фондовом рынке. Использованные статистические данные описывают эволюцию цен акций известной компании. В результате выполнения вычислительных экспериментов установлено, что качество краткосрочных прогнозов волатильности и самого нелинейного нестационарного процесса существенно улучшаются благодаря оптимальной фильтрации данных и рациональной структуре модели. Применение построенной комплексной модели с использованием байесовской сети дало возможность усовершенствовать вероятностное оценивание прогнозов при выполнении торговых операций на фондовом рынке.

Выводы. Оценивание прогнозов нелинейных нестационарных процессов – актуальная задача, которая может быть решена различными методами. Высокоэффективным оказался предложенный вероятностно-статистический метод для оценивания вероятностных прогнозов при выполнении торговых операций на фондовом рынке акций, а поэтому в дальнейшем перспективным будет расширение и усовершенствование его использования.

Ключевые слова: байесовский анализ данных; нелинейные нестационарные процессы; фильтр Калмана; сеть Байеса; регрессионная модель; комбинирование прогнозов.

L.B. Levenchuk, P.I. Bidyuk

BAYESIAN DATA ANALYSIS IN MODELING AND FORECASTING NONLINEAR NONSTATIONARY PROCESSES

Background. Nonlinear nonstationary processes that are available in various spheres of human activity are characterized by numerous uncertainties, fuzziness, incompleteness and low precision data. To perform forecasting of such processes it is necessary to carry out correctly the data processing that is why the problem of development and practical use of the new processing methods is very urgent.

The methods should provide a possibility for performing high quality input data processing aiming to quality modeling and forecasting the processes under study.

Objective. A short review of the Bayesian data analysis methods will be provided and an original methodology for identification and taking into consideration of possible uncertainties that are available in the problems of modeling and forecasting developed. And a new combined probabilistic and statistical model will be proposed for modeling and forecasting nonlinear nonstationary processes.

Methods. A combined implementation methodology has been proposed that includes the following: Bayesian data processing technics, an optimal filter for preliminary data processing, linear and nonlinear regression for formal description and forecasting conditional variance and probabilistic model in the form of Bayesian network to forecast nonlinear nonstationary processes.

Results. The proposed modeling method was tested on the problem of forecast estimation using the financial market data. The statistical data hired describe evolution of stock prices for well-known company. The computational experiments performed showed that quality of the short-term forecasts for volatility and the nonlinear nonstationary financial process itself are improved substantially thanks to the optimal filtering procedure and rational model structure selection. Application of the complex model that uses Bayesian network provided a possibility for improvement of probabilistic forecasts used for performing trade operations at the stock market.

Conclusions. Forecasts estimation for nonlinear nonstationary processes is an urgent problem that can be solved in various ways. The proposed probabilistic and statistical method for estimating probabilistic forecasts used for performing trade operations at the stock market generated high quality results and will be extended and improved in the future.

Keywords: Bayesian data analysis; nonlinear nonstationary processes; Kalman filter; Bayesian network; regression model; forecasts combining.

Рекомендована Радою
Інституту прикладного системного аналізу
КПІ ім. Ігоря Сікорського

Надійшла до редакції
20 грудня 2019 року

Прийнята до публікації
25 червня 2020 року