5. European Statistics Code of Practice. (rev. ed. 2017). ec.europa.eu. Retrieved October 30, 2023 from. URL: https://ec.europa.eu/eurostat/web/quality/european-quality-standards/european-statistics-code-of-practice

БАЙЄСІВСЬКІ НЕПАРАМЕТРИЧНІ МОДЕЛІ ДЛЯ ІНТЕГРАЦІЇ БАГАТОВИМІРНИХ І НЕСТРУКТУРОВАНИХ ДЖЕРЕЛ ДАНИХ У ПРОГНОЗНИЙ ВИСНОВОК

Кобус Олена Сергіївна,
кандидат фізико-математичних наук, доцент
завідувач кафедри технологій захисту
кіберпростору центру кібербезпеки;
Бондаренко Степан Юрійович,
фахівець кафедри технологій захисту кіберпростору центру кібербезпеки;
Національна академія СБ України

Байєсовські непараметричні моделі (англ. BNP) стали інструментом у розробці гнучких структур, які дозволяють інтегрувати багатовимірні та неструктуровані дані в надійні системи прогнозування, особливо за умов, коли звичайні параметричні припущення можуть бути невдалими [3, с. 282]. Такі моделі не передбачають фіксований простір параметрів, а натомість застосовують нескінченновимірний підхід, що дозволяє їм адаптуватися до складності різноманітних ландшафтів даних. Зокрема, застосування моделей BNP для інтеграції різнорідних джерел даних використовує ймовірнісні процеси, особливо процес Діріхле та процеси Гаусса, таким чином полегшуючи моделювання залежностей у різнорідних областях без нав'язування суворих структурних припущень. Ця адаптивність виявляється надзвичайно важливою в областях, де дані за своєю суттю є багатовимірними та охоплюють різні формати, такі як текстова, числова, графічна або категорійна інформація, яка часто протидіє звичайним методам агрегації.

Основна проблема у розгортанні моделей BNP для інтеграції полягає в узгодженні різних форм неструктурованих даних у єдину аналітичну структуру. Завдяки гнучким попереднім і наступним оновленням BNP забезпечує динамічне розміщення даних у міру їх розвитку, часто за допомогою моделей суміші, які враховують приховані структури в даних. Наприклад, процес Діріхле дозволяє вводити в модель нові компоненти зі збільшенням складності даних, уможливлюючи більш детальний аналіз у просторах великої розмірності, де базові закономірності можуть бути приховані в рамках традиційних скінченно-компонентних моделей. Прогнозний висновок у цій структурі значною мірою покладається на створення ієрархічних байєсівських моделей, які дозволяють передавати навчання, ефективно використовуючи інформацію з одного набору

даних для інформування прогнозів щодо іншого, таким чином максимізуючи корисність доступної інформації з різних джерел.

BNP моделі, зокрема ті, що використовують процес Діріхле (англ. DP) і процеси Гаусса (англ. GP), є ключовими інструментами для інтеграції багатовимірних і неструктурованих джерел даних у узгоджені системи прогнозування [1, с. 7]. У цьому контексті моделі BNP вирізняються своєю гнучкістю, оскільки вони не передбачають фіксованих розмірів параметрів і можуть адаптуватися до дедалі складніших наборів даних. Їх здатність структурувати різноманітні та неоднорідні дані – від тексту та зображень до категорійних і числових типів – здебільшого походить від базових стохастичних процесів, які полегшують імовірнісне моделювання без попереднього визначення кількості кластерів даних або прихованих функцій у наборі даних.

Процес Діріхле є центральним для моделей BNP, оскільки він підтримує структуру моделювання суміші, де дані можуть природним чином групуватися в шаблони, коли додається більше спостережень. У програмах, що включають багатовимірні або складні дані, DP забезпечує адаптивну кластеризацію шляхом призначення ймовірностей для нескінченної кількості можливих розділів даних. Ця гнучкість ідеальна для сценаріїв, де кількість базових класів невідома або де з часом очікується поява нових, невідомих типів даних. За допомогою DP кластери даних виникають органічно на основі спостережуваних закономірностей, і через властивість «багатий-стає багатшим» кластери, що представляють домінуючі характеристики даних, мають тенденцію до зростання, оскільки більше даних зміцнює їх структуру. Ця можливість має важливе значення під час інтеграції неструктурованих джерел даних, таких як текст, зображення та сенсорні дані, оскільки вона дозволяє розвиватися у розумінні моделей і зв’язків у цих різноманітних вхідних даних.

З іншого боку, процеси Гауса служать потужними інструментами для моделювання неперервних функцій у складних просторах великої розмірності. У налаштуваннях прогнозування GP пропонує можливості нелінійної регресії, які фіксують кореляції між вимірами даних без нав’язування жорстких лінійних структур. На відміну від процесів Діріхле, які зосереджені на дискретній кластеризації, процеси Гауса працюють над безперервними доменами, що робить їх ідеальними для додатків, де потрібні плавні варіації або передбачення в безперервних просторах, наприклад, аналіз часових рядів, просторове моделювання або дані з часовими залежностями. У контексті неструктурованих даних GP є дуже ефективним для додатків, що включають злиття датчиків, де різні типи потоків даних у реальному часі необхідно інтегрувати в узгоджені прогнози, або в аналізі тексту, де приховані семантичні зв’язки фіксуються у великих корпусах, що розвиваються.

Поєднання структур DP і GP у моделях BNP дозволяє використовувати гібридний підхід, коли неструктуровані дані можна спочатку кластеризувати на основі базових шаблонів, а потім кожен кластер можна проаналізувати за допомогою наближення безперервної функції. Цей гібридний підхід особливо актуальний для багатовимірних джерел даних з категоріальними та

безперервними компонентами, такими як біомедичні дані (наприклад, записи пацієнтів, що поєднують категоричні класифікації захворювань із безперервними біомаркерами), фінансове прогнозування (наприклад, дискретні ринкові події з постійними змінами цін) та аналіз соціальних медіа (наприклад, кластери тем із постійними варіаціями настроїв). Моделюючи як дискретні, так і безперервні структури в даних, моделі BNP дозволяють точніше прогнозувати, що враховує як різкі категоріальні зсуви, так і плавні функціональні залежності.

На практиці використання процесів Діріхле та Гаусса потребує інтенсивних обчислювальних методів, таких як ланцюг Маркова Монте-Карло (MCMC) і варіаційний висновок для наближення апостеріорних розподілів і досягнення логічних висновків. Для процесів Діріхле методи MCMC дозволяють ітераційну кластеризацію, яка динамічно адаптується до нових даних, тоді як у гаусових процесах варіаційний висновок прискорює апостеріорні обчислення у просторах великої розмірності. Ці методи дозволяють ефективно масштабувати моделі BNP навіть із величезними багатовимірними наборами даних, які було б неможливо проаналізувати за допомогою точних обчислень. Крім того, завдяки прогресу в паралелізованих обчисленнях і структурах імовірнісного програмування, моделі BNP, що використовують DP і GP, стають все більш доступними для програм реального часу, таких як виявлення аномалій у кібербезпеці, персоналізовані системи рекомендацій і адаптивне прийняття рішень в автономних системах.

Підсумовуючи, байєсівські непараметричні моделі, що базуються на процесах Діріхле та Гауса, пропонують надійне рішення для інтеграції різноманітних, багатовимірних та неструктурованих джерел даних. Завдяки адаптивній кластеризації DP і безперервному функціональному моделюванню GP, структури BNP дозволяють робити прогнозні висновки, які керуються даними та добре адаптуються до нової інформації. Ця гнучкість має вирішальне значення в епоху, коли джерела даних стають дедалі неоднорідніші, неструктурованими та багатовимірними, що дозволяє моделям BNP служити основними інструментами в галузях, які вимагають всеосяжного прогнозного розуміння динамічних багатогранних наборів даних.

У практичних застосуваннях моделі BNP для інтеграції даних обслуговують різні сфери, включаючи біоінформатику, обробку природної мови та фінансове прогнозування. У біоінформатиці, наприклад, інтеграція генетичних, фенотипових і екологічних даних досягається за допомогою підходів BNP, які пристосовуються до складних взаємозв’язків, притаманних біологічним системам без попередньо визначених розподілів. У фінансах BNP підтримує прогнозування ризиків шляхом синтезу історичних даних про транзакції, ринкових тенденцій і неструктурованих даних новин, забезпечуючи адаптивну оцінку ризиків, яка враховує події в реальному часі. Основна мета тут полягає в створенні прогнозних моделей, які не тільки відображають поточні стани, але й можуть узагальнювати в нових умовах шляхом динамічного оновлення вірувань у міру включення нових даних.

Залежність моделі BNP від імовірнісного висновку також полегшує кількісну оцінку невизначеності в прогнозах. Уникаючи жорстких структурних залежностей, ці моделі забезпечують ширше ймовірнісне розуміння потенційних результатів, що є особливо цінним у сферах прийняття рішень із високими ставками. Байєсівський апостеріорний розподіл дозволяє аналітикам оцінювати мінливість і достовірність своїх прогнозів, таким чином пропонуючи зацікавленим сторонам уявлення як про ймовірні результати, так і про можливі відхилення. Це байєсівське оновлення має вирішальне значення при роботі з невизначеними або неповними джерелами даних, що є типовим у багатовимірних контекстах, де не всі змінні спостерігаються одночасно або в однаковій мірі.

Одним із найпотужніших атрибутів моделей BNP є їх здатність обробляти неструктуровані типи даних, зокрема текстові дані. Використовуючи, наприклад, процеси Гауса, моделі BNP можуть інтерпретувати семантичні нюанси в текстових даних, сприяючи інтеграції таких даних у кількісні прогнозні рамки. Це досягається за допомогою методів вбудовування, які перетворюють текст на числове представлення, яке потім використовується разом з іншими джерелами даних для отримання комплексної інформації, яку інакше було б складно отримати. Ці методи стають дедалі актуальнішими в областях, де текстові дані мають високу інформаційну цінність, наприклад, аналіз відгуків клієнтів або огляди юридичних документів, і їх необхідно узгоджувати з числовими та категоріальними даними для повного прогнозування.

З обчислювальної точки зору розгортання моделей BNP для інтеграції багатовимірних даних є складним завданням, але це стало здійсненним завдяки прогресу в методах вибірки за ланцюгом Маркова Монте-Карло (MCMC) і методів варіаційного висновку. Ці методи полегшують оцінку апостеріорних розподілів у складних просторах даних, забезпечуючи приблизну, але надійну структуру для висновків у контекстах, де точні рішення є непомірно обчислювальними. Методи MCMC, зокрема, дозволяють здійснювати ітеративну вибірку апостеріорних розподілів, поступово зближуючись до прогнозної моделі, яка відображає всю багатовимірну складність даних. Варіаційні методи пропонують швидшу альтернативу, наближаючи апостеріорні розподіли за допомогою методів оптимізації, які особливо корисні в програмах реального часу, де швидкі оновлення є критичними.

Байєсовські непараметричні моделі (BNP), особливо в рамках процесів Діріхле та Гаусса, стають все більш важливими в практичних застосуваннях, які вимагають інтеграції багатовимірних і неструктурованих джерел даних у прогнозну аналітику. Ці моделі відрізняються своєю здатністю працювати без фіксованого простору параметрів, що дозволяє їм гнучко включати різноманітні набори даних, такі як текст, зображення або сенсорні введення, які не мають попередньо визначених структур або категоризацій. Ця адаптивність має вирішальне значення в сучасному контексті, де дані великі й різноманітні, містять як структуровані атрибути (наприклад, категоричні мітки чи числові

значення), так і неструктуровані елементи (наприклад, текстові розповіді чи складну візуальну інформацію).

Однією з ключових переваг моделей BNP в інтеграції неструктурованих джерел даних є їх здатність до адаптивної кластеризації, особливо при реалізації за допомогою процесів Діріхле (DP). На відміну від традиційних моделей, які ґрунтуються на жорстких припущеннях щодо розподілу даних або фіксованих чисел кластерів, DP забезпечує імовірнісний підхід, коли кластери виникають органічно. Це важливо для багатовимірних даних, де кластери не є ні очевидними, ні узгодженими між наборами даних. Наприклад, у сучасній фінансовій аналітиці моделі BNP можуть аналізувати різноманітні дані, такі як ринкові події, текстові звіти та кількісні показники, динамічно групуючи дані на основі статистичних зв'язків, а не довільних критеріїв. Отже, прогностичні висновки, отримані в результаті такої кластеризації, не обмежені початковими припущеннями, що дозволяє високочутливу аналітику, яка підходить для наборів даних, що швидко розвиваються.

GP, з іншого боку, пропонують структуру нелінійної регресії, високоефективну в обробці безперервних структур даних і залежностей між вимірами, що є неоціненним для неструктурованих наборів даних, які вимагають плавної оцінки функції безперервних змінних. Розглядаючи дані як реалізацію базових прихованих функцій, моделі GP допомагають виявити приховані зв'язки, що робить їх особливо придатними для програм, що включають просторові або часові залежності. На практиці процеси Гауса широко застосовуються в програмах, заснованих на навколишньому середовищі та датчиках, де дані в режимі реального часу, такі як показники температури, просторові координати або метрики тиску, постійно інтегруються для генерування чуйної, прогнозованої інформації. В автономних системах, наприклад, це безперервне функціональне моделювання має ключове значення для внесення швидких коригувань на основі вхідних даних у реальному часі, створення чутливої системи, яка може передбачати змінні середовища та пристосовуватися до них.

Практичне застосування моделей BNP у сучасних середовищах з великою кількістю даних також залежить від прогресу в обчислювальних техніках, зокрема у варіаційному висновку та MCMC. Варіаційний висновок прискорює конвергенцію в моделюванні гаусового процесу, особливо при роботі з великими масивами даних великої розмірності, тоді як методи MCMC дозволяють ітераційну кластеризацію, яка адаптується в реальному часі в міру включення нових даних. На практиці ці обчислювальні методи є ключовими для застосовності моделей BNP у реальному часі, оскільки вони дозволяють безперервно генерувати прогнозні висновки у високошвидкісних середовищах, таких як моніторинг безпеки мережі, де великі обсяги різноманітних потоків даних інтегруються та оцінюються на ідентифікація загроз і виявлення аномалій.

Критичним компонентом підходу BNP є його стійкість до переобладнання, що часто є проблемою при роботі з неструктурованими даними, які можуть містити шум або викиди. Імовірнісний характер байєсівських моделей гарантує,

що зроблені висновки базуються на оцінках правдоподібності, а не на детермінованих класифікаціях, що є дуже цінним у таких сферах, як персоналізована медицина, де дані пацієнтів можуть містити унікальні, неповторювані змінні. У медичній діагностиці моделі BNP можуть інтегрувати історію пацієнтів, генетичні дані та показники здоров'я в реальному часі в узгоджений аналіз, який залишається гнучким для унікальних варіацій пацієнтів, водночас генеруючи узагальнені прогнозні ідеї.

Крім того, моделі BNP підтримують прийняття рішень у невизначених середовищах, де неможливо гарантувати повноту чи якість даних [2, с. 62]. Наприклад, у стратегічних програмах, таких як оборона чи аналіз розвідки, інформація часто фрагментована або неповна, що вимагає моделей, які можуть синтезувати багатовимірні, неструктуровані дані з різних джерел, таких як супутникові зображення, соціальні мережі та польові звіти. Здатність BNP працювати з неповними структурами даних, використовуючи передбачуваний розподіл для визначення відсутніх точок даних на основі спостережуваних шаблонів, забезпечує більш стійке та обґрунтоване прийняття рішень, навіть за невизначених або суперечливих умов.

Таким чином, байєсівські непараметричні моделі забезпечують складну адаптовану структуру для інтеграції багатовимірних і неструктурованих джерел даних у прогнозні висновки. Їхня здатність моделювати як дискретні, так і безперервні латентні структури без попередньо визначених припущень дозволяє їм динамічно адаптуватися до нової інформації, що є важливою ознакою в сучасних умовах, де дані об'ємні, різноманітні та постійно змінюються. Завдяки комбінованому застосуванню процесів Діріхле та Гауса моделі BNP фіксують складні приховані зв'язки всередині даних, розвиваючи передбачувані можливості в сферах, які вимагають гнучкої аналітики в реальному часі в різних джерелах даних. Практичному розгортанню моделей BNP сьогодні сприяють передові обчислювальні методи, які забезпечують масштабованість і адаптивність у реальному часі, що робить BNP незамінним підходом у середовищах, де складність і різноманітність даних представляють постійні проблеми.

Підсумовуючи, байєсівські непараметричні моделі пропонують складний підхід до інтеграції багатовимірних і неструктурованих джерел даних, надаючи прогнозні висновки, які є одночасно гнучкими та адаптованими до мінливих інформаційних ландшафтів. Завдяки застосуванню нескінченновимірних процесів моделі BNP забезпечують нюансовану імовірнісну структуру, яка поважає притаманну складність різнорідних даних. Ця адаптивність робить моделі BNP особливо актуальними в сучасній прогностичній аналітиці, де дані з різних областей повинні бути синтезовані в узгоджені, практичні ідеї. Оскільки обчислювальні методи продовжують розвиватися, очікується, що здійсненність і застосування моделей BNP в інтеграції даних і прогнозному моделюванні розширюватимуться, пропонуючи глибші ймовірнісніше обґрунтовані уявлення про дедалі складніші середовища даних. Роль байєсівських непараметричних

моделей, таким чином, виступає як трансформаційна та незамінна у розвитку можливостей прийняття рішень на основі даних.

Список використаних джерел

1. Gershman S. J., Blei D. M. A tutorial on Bayesian nonparametric models. Journal of Mathematical Psychology. 2012. Т. 56, № 1. С. 1–12. URL: https://doi.org/10.1016/j.jmp.2011.08.004 (дата звернення: 11.11.2024).

2. Lee D., Pan R. A nonparametric Bayesian network approach to assessing system reliability at early design stages. Reliability Engineering & System Safety. 2018. Т. 171. С. 57–66. URL: https://doi.org/10.1016/j.ress.2017.11.009 (дата звернення: 11.11.2024).

3. Müller P., Mitra R. Bayesian Nonparametric Inference – Why and How. Bayesian Analysis. 2013. Т. 8, № 2. С. 269–302. URL: https://doi.org/10.1214/13-ba811 (дата звернення: 11.11.2024).