ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ІМПУТАЦІЇ ДАНИХ ЗМІШАНОЇ ПРИРОДИ В ЗАДАЧАХ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ

УДК 303.7:004.6

СЛАБЧЕНКО Олеся Олегівна
пошукувач, кафедра комп'ютерних та інформаційних систем,
Кременчуцький національний університет імені Михайла Остроградського.
Наукові інтереси: інтелектуальний аналіз даних, соціально-мережевий аналіз.
E-mail: slabchenko.olesia@gmail.com.

ВСТУП

Незадовільна якість первинних даних внаслідок наявності пропущених значень є однією з основних проблем при аналізі реальних даних численних предметних областей. Розвиток Інтернет-платформ обумовлює накопичення значної кількості доступної для обробки інформації, яка використовується прикладними методами інтелектуального і соціально-мережевого аналізу. Однак, дослідження у цій сфері підтверджують, що основною проблемою даних такого типу залишається значна частка пропущених значень, яка ускладнюється змішаною природою показників. Для застосування спеціалізованих методів і алгоритмів аналізу даних необхідно, щоб вхідні показники задовольняли умові комплектності, тобто не містили пропусків. Вирішення проблеми некомплектних даних є важливим підготовчим етапом, що передує процесу моделювання, і, згідно з методологією реалізації Data mining проектів CRISP-DM [1], повинне відбуватися на етапі підготовки (Data Preparation) даних. Вибір методу обробки пропущених значень обумовлюється рядом чинників, серед яких ключову роль відіграють механізм утворення пропусків [2] і тип показників. Неправильне опрацювання некомплектних даних може призводити до отримання значних похибок і зміщених результатів аналізу [3]. Враховуючи вимоги й обмеження щодо застосування методів обробки пропусків, а також особливості даних із онлайн-платформ, найбільш перспективним і ефективним є застосування методу імпутації (відновлення) пропущених значень [4]. Імпутація – це процедура оцінки невідомих значень показників на основі наявних даних, результатом якої є отримання комплектної множини даних, що може оброблюватися далі необхідними методами інтелектуального аналізу. Дослідження у сфері імпутації некомплектних показників із онлайн-платформ показують, що хоча й існує ряд методів обробки даних подібного типу, однак відсутні технології, що дозволяють автоматизувати процедуру їх відновлення. Тому актуальною є задача розробки інформаційної технології імпутації пропущених даних, яка б дозволила автоматизувати процедуру їх підготовки до подальшого аналізу. Враховуючи вищесказане,

метою роботи є підвищення якості первинних даних із онлайн-платформ на етапі їх попередньої обробки шляхом розробки інформаційної технології відновлення пропущених даних.

ВИКЛАДЕННЯ ОСНОВНОГО МАТЕРІАЛУ

Розглянемо детальніше методологію реалізації проектів інтелектуального аналізу даних CRISP-DM, яка протягом останнього десятиліття незмінно залишається найбільш популярною серед аналітиків [5]. Вона передбачає розбиття процесу аналізу на шість основних етапів (рис. 1): розуміння бізнесу, розуміння даних, попередня обробка даних, моделювання, оцінка й розгортання [6].

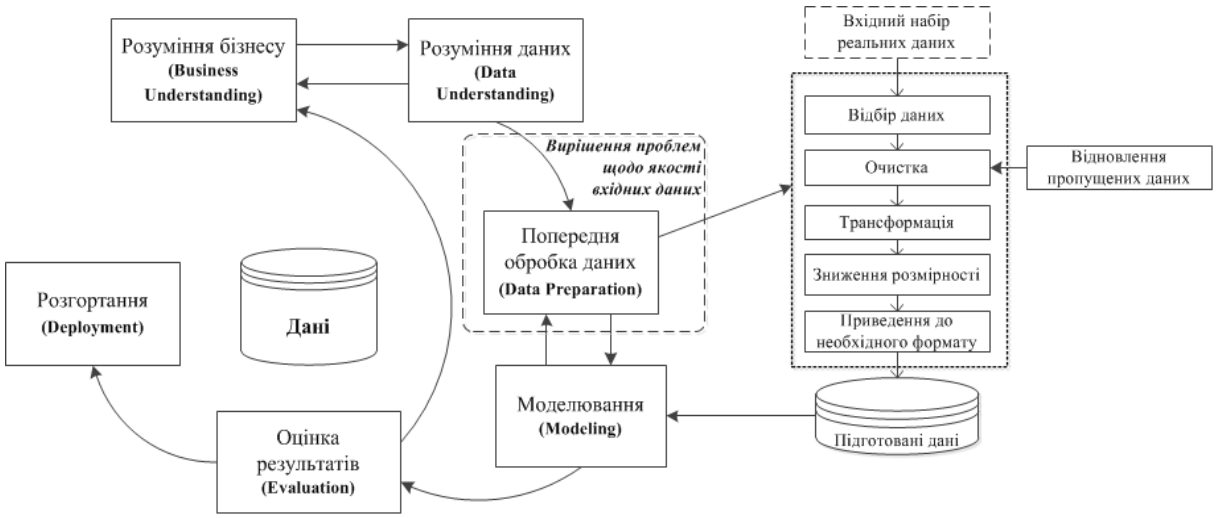

Рисунок 1 – Методологія аналізу даних CRISP-DM і етапи, на яких відбувається вирішення проблем щодо якості даних

Як видно з рисунку, власне процесу побудови моделей передує аж три етапи, на яких основна увага приділяється розумінню структури, цінності даних і способам їх підготовки до аналізу, оскільки від цього залежить достовірність і стійкість результатів моделювання. Етап попередньої обробки включає відбір, опис, дослідження якості даних, а також передбачає відбір показників для подальшого аналізу, вирішення проблем щодо їх якості, очистку, трансформацію, зниження розмірності й приведення з різнорідних форматів до єдиного, зручного для аналізу. Підвищення якості первинних даних відбувається саме на цьому етапі, однак ускладнюється слабкою структурованістю і значною кількістю унікальних значень атрибутів [4]. Впровадження інформаційної технології відновлення пропущених значень повинне відбуватися на етапі підготовки даних до моделювання і дозволить автоматизувати процедуру підвищення якості вихідних показників шляхом імпутації некомплектних значень. У загальному вигляді процес обробки первинних некомплектних даних із атрибутів онлайн-платформ і приведення їх до комплектного виду можна представити наступним чином (рис. 2):

Рисунок 2 – Процес відновлення пропущених даних

Інформаційна технологія (ІТ) – процес, який використовує сукупність методів і засобів реалізації операцій збору, реєстрації, передачі, накопичення й обробки інформації на базі програмно-апаратного забезпечення для вирішення управлінських задач економічного об'єкту [7]. У даному випадку основною метою побудови ІТ відновлення пропущених даних є отримання інформації нової якості шляхом поетапної обробки некомплектної множини вхідних даних із застосуванням методів і моделей імпутації, на основі якої можна виконувати процес моделювання і забезпечувати функціонування систем підтримки прийняття рішень (СППР). Розроблена інформаційна технологія (рис. 3) включає наступні складові: вхідні дані, комплекс математичних методів, моделей і ансамблів для їх аналізу, а також інформаційні процеси (ІП) і структуру взаємодії між ними. Визначено ряд наступних інформаційних процесів: підготовка даних до моделювання, кластеризація даних, визначення набору синтезованих моделей, формування навчальної і тестової множин даних, синтез моделей імпутації, синтез ансамблів моделей відновлення даних, імпутація пропущених даних, формування комплектної множини.

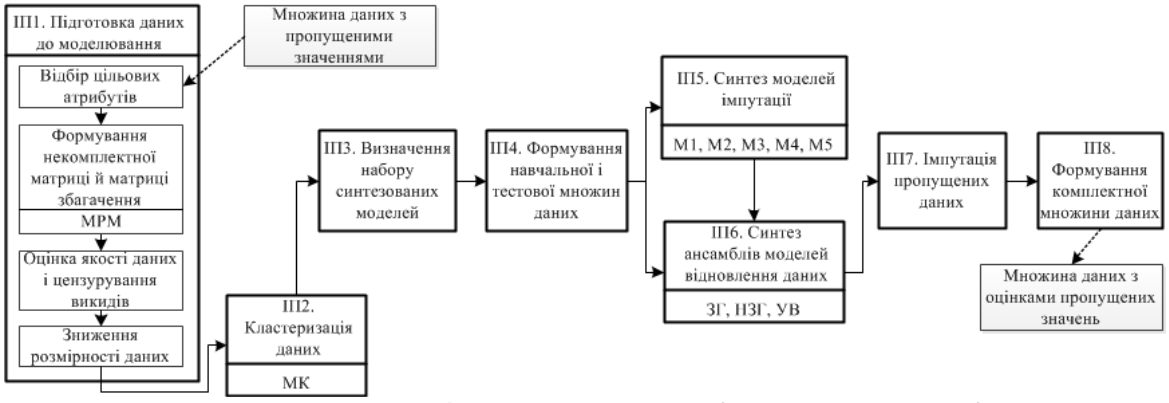

Рисунок 3 – Загальна схема інформаційної технології відновлення пропущених даних

Розглянемо детальніше інформаційні процеси розробленої ІТ.

ІП1. Підготовка даних до моделювання. Нехай Z – матриця первинних даних розмірністю $m \times n$, елементи якої описують n об'єктів у просторі m кількісно-якісних атрибутів. Набір подібних атрибутів можна розділити на 2 групи: ті, що потенційно можуть містити пропущені значення (звичайно анкетні дані) й такі, що завжди комплектні (службові дані, які зберігаються і використовуються системою). На основі інформації щодо комплектності наявних даних інформаційний процес підготовки даних до моделювання складається з наступних етапів:

1.1 Відбір k_1 цільових некомплектних показників, які потребують імпутації.

1.2 Застосування методу формування розширеної матриці атрибутів (МРМ), який включає формування некомплектної матриці X_1 розмірністю $k_1 \times n$ на основі атрибутів з пропущеними значеннями і матриці збагачення X_2 розмірністю $\mathrm{k_2 \times n}$ на основі k_2 завжди комплектних атрибутів. У результаті об'єднання $X_1 + X_2$ отримують розширену матрицю атрибутів X розмірністю $\mathrm{k \times n}$, де $k = k_1 + k_2$ [8].

1.3 Оцінка якості даних розширеної матриці атрибутів X, пошук викидів й екстремальних значень і їх цензурування.

1.4 Зниження розмірності матриці X методом факторного аналізу.

Кінцевим результатом ІП1 є цензурована розширена матриця атрибутів X із зниженою розмірністю, підготована до подальшого аналізу.

ІП2. Кластеризація даних. Для вирішення проблеми значної кількості унікальних значень атрибутів, що ускладнює обробку даних із онлайн-платформ [4, 9], застосовується метод на основі процедури попередньої кластеризації (МК). Його ідея полягає в пошуку гомогенних груп об'єктів, утворених на основі схожих дескриптивних апріорно-апостеріорних даних. Виявлення подібних кластерів дозволяє диференціювати набори даних з однотипними значеннями і, відповідно, знизити кількість унікальних значень атрибутів всередині них порівняно з некластеризованою матрицею.

Кінцевим результатом ІП2 є s множин даних кластерів $X_q (q = \overline{1,s})$, де s – кількість знайдених кластерів, всередині яких можуть застосовуватися подальші методи аналізу.

ІП3. Визначення набору синтезованих моделей. Атрибути, які містять некомплектні дані, можуть мати різну природу (числову або номінальну). Оскільки серед моделей імпутації, які розглядаються в роботі далі, не всі є універсальними і здатними до одночасної обробки даних різного типу, при аналізі конкретних некомплектних множин X_q відбувається відбір математичних моделей, які будуть застосовуватись у процесі відновлення пропущених значень.

Результатом ІП3 є деяка множина моделей $M = \{ M_1 M_2, \ldots, M_c \}$, де c – кількість відібраних моделей імпутації, на основі яких відбуватиметься подальший процес відновлення пропущених значень.

ІП4. Формування навчальної і тестової множин. Для застосування обраних моделей на основі інформації щодо наявності/відсутності значень атрибутів, які підлягають імпутації, необхідно сформувати навчальні й тестові множини. Нехай $X = \left(X_1, X_2, \ldots, X_k \right)$ – некомплектна матриця даних, що містить k атрибутів, а вектору X_i $(i = \overline{1,k})$ відповідає вектор значень атрибуту з номером i. Позначимо через m кількість пропущених значень у довільному атрибуті з номером i. Тоді при формуванні навчальної множини для відновлення некомплектних даних цього атрибуту відбувається виключення m рядків матриці X, яким відповідають пропуски у векторі значень відповідного атрибуту $\mathrm{X_i}$. У результаті отримують множину $\mathrm{X^{train}}$, в якій всі значення некомплектного атрибуту i-го відомі. До тестової множини включаються m рядків матриці X, які містять тільки некомплектні дані атрибуту i-го: $X^{test} = X \setminus X^{train}$.

Кінцевим результатом ІП4 є навчальна множина $\mathrm{X^{train}}$, яка використовується для налаштування моделей імпутації, і тестова

множина X^{test}, в якій здійснюється відновлення пропущених значень.

ІП5. Синтез моделей імпутації. Запропонована інформаційна технологія основана на застосуванні п'яти моделей імпутації: на основі асоціативних правил (AR/M1) [4, 10], випадкового лісу (RF/M2) [8, 11], машини опорних векторів (SVM/M3), нейронної мережі (архітектура – багатошаровий перцептрон з двома прихованими шарами) (ANN/M4) та EM-алгоритму (EM/M5) [4]. Моделі AR і RF є універсальними й дозволяють одночасну обробку даних різного типу, SVM і ANN використовують номінальні атрибути лише в якості міток класів, а EM не включає у процес аналізу нечислові показники, що враховується на попередньому етапі при формуванні навчальних і тестових множин. Кожну з математичних моделей можна представити у вигляді кортежу $M < p_1, p_2, \ldots, p_q >$, у якому параметри $p_j (\, j = \overline{1,q})$ визначають її роботу. ІП5 включає налаштування ряду початкових параметрів множини моделей імпутації, сформованої у результаті ІП3, і їх навчання на даних множини X^{train}.

Кінцевим результатом ІП5 є множина навчених моделей імпутації $M = \{ M_1, M_2, \ldots, M_c \}$ з визначеними параметрами $< p_1, p_2, \ldots, p_q >$.

ІП6. Синтез ансамблів моделей відновлення даних. З метою підвищення якості і стабільності процесу імпутації на основі навчених моделей, отриманих у результаті ІП5, виконується побудова їх ансамблю, оскільки використання моделей різних типів надає класифікатору додаткову гнучкість [12]. У залежності від набору синтезованих моделей, визначеного в результаті ІП3, відбувається комбінація результатів, які видають окремі моделі.

У випадку імпутації номінального атрибуту використовується зважене голосування (ЗГ), оскільки якість роботи моделей при відновленні нечислових даних відрізняється [4]. Для врахування рівня достовірності результатів кожній із n моделей ансамблю присвоюється певна вага $w_i (i = \overline{1, n})$ таким чином, що сума ваг усіх моделей $\sum_{i=1}^{n} w_i = 1$. Вихід кожної з моделей позначається через y_i. Відповідно вихід ансамблю Y формується наступним чином: $Y = y_1 \cdot w_1 + y_2 \cdot w_2 + \cdots + y_i \cdot w_i$, в результаті чого вихід моделі (або моделей), який набрав найбільшу вагу, визначає результуючий вихід ансамблю.

При імпутації числових атрибутів використовуються два інші види комбінування результату: незважене голосування (НЗГ) й незважене усереднення (УВ). При незваженому голосуванні усі моделі мають однакову вагу і результуючий вихід обирається простою більшістю голосів. При незваженому усередненні вихід ансамблю визначається як середнє значення виходів усіх моделей $Y = (y_1 + y_2 + \cdots + y_n) / n$.

Кінцевим результатом ІП6 є сформований ансамбль моделей імпутації з визначеним варіантом комбінування виходу залежно від типу атрибуту, значення якого відновлюються, і множини використовуваних моделей, отриманих у результаті ІП3.

ІП7. Імпутація пропущених значень. На вхід отриманого ансамблю подається множина X^{test}, яка містить некомплектні значення атрибуту, що потребують імпутації. Після цього навчений ансамбль моделей виконує оцінку кожного пропущеного значення на основі комплектних даних тестової множини. Результатом ІП7 є

вектор оцінок пропущених значень Y_k некомплектного атрибуту k.

ІП8. Формування комплектної множини даних. У вхідній некомплектній матриці X відбувається підстановка оцінок пропущених значень атрибуту Y_k, отриманих у результаті застосування ансамблю, замість кожного відповідного пропуску атрибуту k. Кінцевим результатом ІП8 є комплектна множина даних X^{imp} з точковими оцінками значень номінальних атрибутів й інтервальними оцінками значень числових атрибутів з допущенням відхилення в діапазоні Δr, $r \in [0;1]$.

ВИСНОВКИ

Розроблено інформаційну технологію імпутації даних змішаної природи, яка дозволяє підвищувати якість первинних даних у задачах інтелектуального і соціально-мережевого аналізу. Технологія включає комплекс математичних моделей, методів і систему інформаційних процесів, що реалізують процедури отримання, обробки, зберігання і видачі інформації, яка використовується в процесі функціонування ІТ. Метод формування розширеної матриці атрибутів дозволяє доповнити вихідну матрицю неповних даних матрицею збагачення завжди комплектних показників і таким чином підвищити інформативність початкових даних у процесі їх аналізу. Метод попередньої кластеризації даних дозволяє знизити кількість унікальних значень атрибутів і таким чином спростити застосування подальших методів обробки даних. Моделі імпутації на основі асоціативних правил, лісу рішень, машини опорних векторів, нейронної мережі та EM-алгоритму виконують відновлення пропущених значень атрибутів і приводять множину первинних даних до комплектного вигляду, що дозволяє застосовувати до неї необхідні подальші методи й алгоритми аналізу, які вимагають на вхід повні набори даних. Ансамблі моделей імпутації за рахунок використання моделей різних типів дають можливість підвищувати якість і стійкість результатів відновлення даних, що показують дослідження: відхилення значень коректності імпутації пропусків ансамблями від відповідних результатів найкращих моделей знаходиться в межах ±4%, у той час як відхилення результатів менш ефективних одиночних моделей може складати 8–24%. Загалом, розроблена інформаційна технологія дозволяє автоматизувати процедуру відновлення пропущених значень на етапі їх підготовки до аналізу і таким чином підвищити якість первинних даних, від якої залежить достовірність і ефективність подальшого процесу моделювання.

ЛІТЕРАТУРА

1. Azervedo, A. KDD, semma and CRISP-DM: A parallel overview / A. Azervedo, M. F. Santos // Proceedings of IADIS European Conference on Data Mining. Amsterdam, July 24–26, 2008. – IADIS Press. – 2008. – P.182–185.
2. Graham, J. W. Missing Data: Analysis and Design / J. W. Graham. – New York: Springer, 2012. – 324 p.
3. Nakagawa1, S. Missing inaction: the dangers of ignoring missing data / S. Nakagawal, R. P. Freckleton // Trends in Ecology and Evolution. – 2008. – Vol. 23, Iss. 11. – P.592–596.
4. Slabchenko, O. O. Efektyvnist' zastosuvannya metodiv ta alhorytmiv imputatsiyi propushchenykh danykh v zadachakh sotsial'no-merezhevoho analizu /O. O. Slabchenko, V. M. Sydorenko // Visnyk Kremenchuts'koho natsional'noho universytetu imeni Mykhayla Ostrohrads'koho. – 2016. – № 2(97). S.15–26.
5. Piatetsky, G. CRISP-DM, still the top methodology for analytics, data mining, or data science projects [Elektronnij resurs] / Rezhim dostupu: http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html – 2014. – Zagol. z ekranu.

6. Wirth, R. CRISP-DM: Towards a standard process model for data mining / R. Wirth // Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining. Manchester, April 11–13, 2000. – Practical Application Company. – 2000. – P.29–39.
7. Informatsyonnye tekhnolohyy upravlenyya: Ucheb. Posobye dlya vuzov / Pod red. prof. H.A. Tytorenko. – 2-e yzd., dop. – M.: YUNYTY-DANA, 2003. – 439 s.
8. Slabchenko, O. Analysis and synthesis of models on basis of machine learning for missing values imputation from social networks' personal accounts / O. Slabchenko, V. Sydorenko // Visnyk Kremenchuts'koho natsional'noho universytetu imeni Mykhayla Ostrohrads'koho. – 2014. – № 6(88). – S.105–111.
9. Slabchenko, O. The improvement of initial data quality in modeling problems of online communities on the base of combined implementation of segmentation, imputation and data enrichment models / O. Slabchenko, V. Sydorenko // Visnyk Kremenchuts'koho natsional'noho universytetu imeni Mykhayla Ostrohrads'koho. – 2013. – №6(83). S.50–58.
10. Slabchenko, O. An improved algorithm for imputation data from social network accounts with use of association rules / O. Slabchenko, V. Sydorenko, X. Siebert // Zbirnyk materialiv XXI Mizhnarodnoyi naukovo-tekhnichnoyi konferentsiyi studentiv, aspirantiv ta molodykh uchenykh KrNU imeni Mykhayla Ostrohrads'koho «Aktual'ni problemy zhyttyediyal'nosti suspil'stva». 24–25 kvitnya 2014 r. – Kremenchuk: KNU imeni Mykhayla Ostrohrads'koho. – 2014. – S.45–46.
11. Slabchenko, O. Models of imputation missing data from social networks' accounts on basis of decision trees and random forests / O. Slabchenko, V. Sydorenko, X. Siebert // Zbirnyk materialiv XXII Mizhnarodnoyi naukovo-praktychnoyi konferentsiyi studentiv, aspirantiv ta molodykh uchenykh KrNU imeni Mykhayla Ostrohrads'koho «Aktual'ni problemy zhyttyediyal'nosti suspil'stva». 15–16 kvitnya 2015 r. – Kremenchuk: KNU imeni Mykhayla Ostrohrads'koho. – 2015. – S.37–39.
12. Rokach, L. Ensemble-based classifiers / L. Rokach // Artificial Intelligence Review. – 2010. – № 1(33). – P.1–39.

Рецензент: д.т.н., проф. Чорний О.П.
Інститут електромеханіки, енергозбереження і систем управління