УДК 303.7:004.6

ЕФЕКТИВНІСТЬ ЗАСТОСУВАННЯ МЕТОДІВ ТА АЛГОРИТМІВ ІМПУТАЦІЇ ПРОПУЩЕНИХ ДАНИХ В ЗАДАЧАХ СОЦІАЛЬНО-МЕРЕЖЕВОГО АНАЛІЗУ

О. О. Слабченко, В. М. Сидоренко
Кременчуцький національний університет імені Михайла Остроградського
вул. Першотравнева, 20, м. Кременчук, 39600, Україна. E-mail: slabchenko.olesia@gmail.com

Досліджено структуру апріорно-апостеріорних даних із акаунтів соціальних мереж і виявлено дві групи показників: потенційно некомплектні і завжди комплектні. Висунуто гіпотезу щодо впливу матриці комплектних показників на ефективність процесу імпутації і сформульовано поняття розширеної матриці даних. Для вирішення проблеми значної кількості унікальних значень атрибутів запропоновано підхід на основі попередньої кластеризації вихідних даних. Виконано синтез моделей імпутації на основі машинного навчання з використанням кластеризації і розширеної матриці даних. Досліджено ефективність застосування розроблених моделей для даних номінального і числового типу з точки зору оцінки двох різних похибок. Показано, що етап попередньої кластеризації дозволяє підвищувати коректність відновлення пропущених значень різного типу. Виявлено, що використання розширеної матриці даних доцільне лише для номінальних даних. Здійснено відбір оптимальних моделей імпутації даних різного типу для подальшого застосування.

Ключові слова: імпутація, розширена матриця даних, попередня кластеризація, асоціативні правила, випадкові ліси, машини опорних векторів, нейронні мережі, ЕМ-алгоритм.

ЭФФЕКТИВНОСТЬ ПРИМЕНЕНИЯ МЕТОДОВ И АЛГОРИТМОВ ИМПУТАЦИИ ПРОПУЩЕННЫХ ДАННЫХ В ЗАДАЧАХ СОЦИАЛЬНО-СЕТЕВОГО АНАЛИЗА

О. О. Слабченко, В. Н. Сидоренко
Кременчугский национальный университет имени Михаила Остроградского
ул. Первомайская, 20, 39600, г. Кременчуг, Украина. E-mail: slabchenko.olesia@gmail.com

Исследована структура априорно-апостериорных данных из аккаунтов пользователей социальных сетей и выявлены две группы показателей: потенциально некомплектные и всегда комплектные. Выдвинута гипотеза относительно влияния матрицы комплектных показателей на эффективность процесса импутации и сформулировано понятие расширенной матрицы данных. Для решения проблемы значительного количества уникальных значений атрибутов предложен подход на основе предварительной кластеризации исходных данных. Синтезированы модели импутации данных на основе машинного обучения с использованием кластеризации и расширенной матрицы данных. Исследована эффективность применения разработанных моделей для данных номинального и числового типа с точки зрения двух погрешностей. Показано, что этап предварительной кластеризации позволяет повышать корректность восстановления пропущенных значений различного типа. Обнаружено, что использование расширенной матрицы данных целесообразно только для номинальных данных. Осуществлен отбор оптимальных моделей импутации данных различного типа для дальнейшего применения.

Ключевые слова: импутация, расширенная матрица данных, предварительная кластеризация, ассоциативные правила, случайные леса, машины опорных векторов, нейронные сети, ЕМ-алгоритм.

АКТУАЛЬНІСТЬ РОБОТИ. Як показують численні дослідження, наявність пропущених значень у наборах реальних даних є однією із основних проблем при їх обробці [1, 2]. Особливо актуальною дана задача стає при застосуванні алгоритмів аналізу даних, які вимагають подачу комплектних множин на вхід. Аналіз літературних джерел показує, що на сьогодні існує ряд підходів до обробки некомплектних показників. Згідно з класифікацією, запропонованою Літтлом і Рубіном [3], виділяють методи видалення, зважування, на основі моделі породження пропусків і заповнення (імпутації), застосування яких визначається особливістю аналізованих даних. У випадку побудови моделей соціально-мережевого аналізу (Social Network Analysis, SNA), основою для яких є показники з акаунтів користувачів соціальних мереж, для ефективної й адекватної обробки даних із пропущеними значеннями необхідно приймати до уваги їх специфічні особливості, а саме: зв'язну природу і високу ймовірність присутності кореляцій між показниками.

Застосування підходів із групи видалення призводить до зміщення оцінок параметрів і статистичних тестів за присутності значимого кореляційного зв'язку між змінними. Методи зважування також дають незміщену оцінку лише за припущення, що ймовірність виникнення пропущених значень однакова для всіх елементів вибірки. Методи на основі моделі породження пропусків потребують попереднього визначення механізму виникнення некомплектних значень для його подальшого включення в результуючу модель з метою отримання незміщених оцінок. Оскільки при розгляді даних із соціальних мереж механізм виникнення пропущених значень знаходиться поза контролем дослідника, його розподіл є невідомим, що створює труднощі або робить неможливим застосування цієї групи методів. На відміну від перших трьох підходів, які ніяким чином не заповнюють пропущені значення, а лише виконують оцінку спостережуваних при виконанні необхідних умов, методи імпутації дозволяють замінювати пропуски деякими правдоподібними оцінками. Відповідно не відбувається втрати інформації, а якщо наявні дані містять інформацію про характер пропусків, вона може бути використана для отримання кращих оцінок пропущених значень.

Враховуючи ряд обмежень, що накладаються на застосування методів видалення, зважування і на основі моделі породження пропусків, а також негативні наслідки виключення частини доступної інформації

для адекватності й надійності моделювання [4], застосування методів імпутації є найбільш обґрунтованим і перспективним підходом, що підтверджується популярністю і розробленістю даної області.

Метою роботи є підвищення якості вихідних даних із акаунтів користувачів соціальних мереж на етапі їх попередньої обробки шляхом застосування методів імпутації пропущених значень.

МАТЕРІАЛ І РЕЗУЛЬТАТИ ДОСЛІДЖЕНЬ. *Опис атрибутів із акаунтів соціальних мереж. Відбір цільових показників*. Розглянемо структуру даних, які містять акаунти користувачів соціальних мереж, на прикладі сайту «Вконтакте». Дослідження показують, що вони можуть зберігати порядку декількох десятків доступних для аналізу показників різного типу: анкетні дані, текстовий, мультимедійний контент,інформація щодо активності, зв'язків, налаштувань профілю тощо. У рамках соціально-мережевого аналізу подібні дані представляються у вигляді матриці «об'єкт-ознака» – прямокутної таблиці з числом рядків, рівним кількості користувачів, і стовпчиків, що дорівнюють числу атрибутів, які їх описують. Зазвичай, цільовою при застосуванні методів соціально-мережевого аналізу є анкетна інформація (вік, стать, рівень освіти, географічне місцезнаходження, сімейний стан тощо). Однак, значна кількість інших специфічних даних також може бути задіяною в аналізі безпосередньо або після деякої трансформації виходячи зі специфічних вимог і задач досліджуваної прикладної області. Залежно від природи виникнення, характеру інформації, що зберігається в наведених атрибутах профілів, і її динаміки диференціюємо показники на апріорні та апостеріорні. Апріорні дані містять анкетну інформацію про користувачів (вік, стать, освіту, хобі тощо), контактні дані, налаштування акаунта і т. ін. Апостеріорні визначаються активністю відвідування мережі й інтенсивністю та характером взаємодії з іншими користувачами (динаміка створення і видалення зв'язків, завантаження і оперування файлами тощо).

У результаті певних модифікацій та агрегації ряд доступних атрибутів можна привести до апостеріорних показників, які відсутні в явному вигляді, але мають певний смисловий зміст. Така процедура по суті є внутрішнім збагаченням даних за рахунок зміни їх організації. На рис. 1 наведено результуючу структуру апріорно-апостеріорних атрибутів і показників, обчислених на їх основі, що містяться в досліджуваній соціальній мережі.

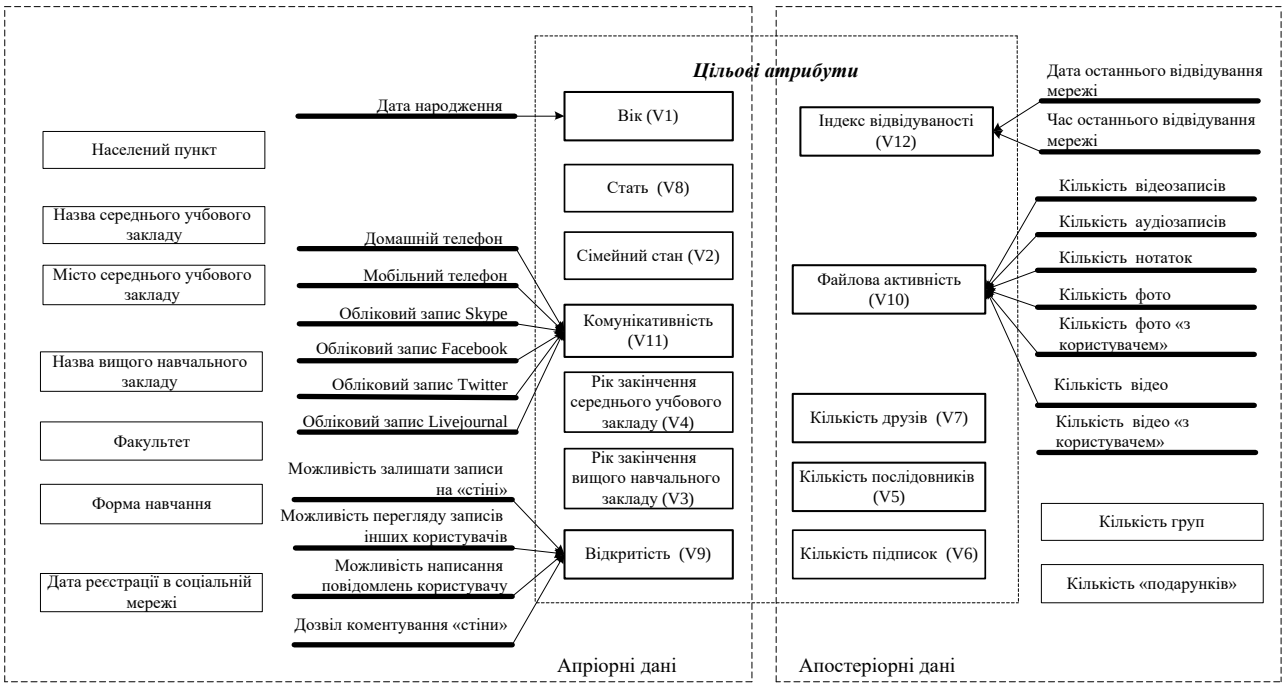

Рисунок 1 – Структура апріорно-апостеріорних показників із акаунтів соціальних мереж

Як видно з рисунку, інформація щодо дати народження трансформується в показник «Вік»; дані щодо часу і дати останнього відвідування мережі приводяться до показника «Індекс відвідування», який характеризує частоту входження на сайт; дані про мультимедійний контент перетворюються на показник «Файлова активність»; інформація щодо налаштувань приватності профілю приводиться до показника «Відкритість»; дані про координати користувача поза досліджуваною соціальною мережею модифікуються в показник «Комунікативність».

Таким чином, після дослідження й трансформації ряду атрибутів із акаунтів соціальної мережі відібрано множину дванадцяти цільових для подальшого аналізу показників, позначених на рисунку як V1–V12.

Поняття розширеної матриці даних. Розглянемо далі структуру обраних показників з точки зору їх комплектності. Для цього сформуємо множину реальних даних користувачів соціальної мережі на основі інформації з їх акаунтів. У табл. 1 наведено дані щодо ступеня комплектності атрибутів, обраних для аналізу у попередньому пункті. Множину показників, що описують користувача, можна розді-

лити на 2 групи: атрибути, які можуть містити пропущені дані (V1–V4: вік, сімейний стан, рік закінчення вищого і середнього навчального закладів), і абсолютно комплектні атрибути (V5–V12):

Таблиця 1 – Комплектність атрибутів

Атрибут	Частка комплектних значень
V1	45,25 %
V2	57,99 %
V3	79,2 %
V4	74,26 %
V5	100 %
V6	100 %
V7	100 %
V8	100 %
V9	100 %
V10	100 %
V11	100 %
V12	100 %

Природа показників другого типу пояснюється тим, що структура організації зберігання інформації в соціальній мережі, окрім даних про самого користувача, передбачає також наявність службової інформації, необхідної для функціонування мережі (наприклад, показник «Відкритість» обчислюється на основі параметрів приватності профілю, «Індекс відвідуваності» – на основі даних щодо останнього користування мережею). Ряд інших показників зв'язків, таких як кількість друзів, послідовників, підписок, лічильників файлів не може містити пропущених значень за своїм визначенням, оскільки також використовується мережею й іншими користувачами.

Таким чином, результати досліджень показують, що ряд атрибутів акаунтів містить пропущені значення. Множину таких даних будемо називати некомплектною матрицею. Поряд із ними соціальна мережа також містить ряд атрибутів, які є завжди заповненими. Матрицю комплектних даних будемо називати матрицею збагачення, оскільки показники, які містяться в ній, не лише представляють інтерес для аналізу, а й можуть містити інформацію, яка доповнює некомплектну матрицю і може бути використана у процесі імпутації пропущених даних. Відповідно матрицю, отриману в результаті об'єднання некомплектної матриці і матриці збагачення, будемо називати розширеною матрицею даних, яка в даному випадку включає чотири потенційно некомплектних показники (V1–V4) і вісім завжди заповнених (V5–V12).

Інтерес для подальшого дослідження представляє гіпотеза щодо впливу матриці збагачення на якість результатів імпутації. Далі у роботі розглядається два підходи до процесу відновлення пропущених даних: побудова і навчання моделей суто на даних некомплектної матриці і використання розширеної матриці, до складу якої входить також матриця збагачення.

Підхід до імпутації на основі попередньої кластеризації. Проведений аналіз структури атрибутів, що містяться в акаунтах соціальної мережі, дозволяє також виявити проблеми, що ускладнюють їх подальшу обробку. Як видно із рис. 1, цільові атрибути неоднорідні за своїм типом і розділяються на числові (відносні та порядкові), номінальні й дихотомічні. У першу чергу, це обумовлює необхідність вибору алгоритму, який працює з даними змішаної природи, для наступної процедури їх аналізу. Виходячи з результатів досліджень, іншою проблемою, характерною для множин даних із соціальних мереж, є наявність атрибутів із значною кількістю унікальних значень [5]. Для вирішення останньої у роботі описується підхід на основі попередньої кластеризації, вперше запропонований авторами у покращеному алгоритмі імпутації на основі асоціативних правил [6]. Його ідея полягає в тому, що припускається існування деяких гомогенних груп користувачів на основі їх апріорних даних і поведінки в мережі, які можуть утворювати природні сегменти. Знаходження таких кластерів і застосування алгоритмів аналізу всередині кожного з них дозволить знизити кількість унікальних значень показників, що розглядаються, порівняно з некластеризованою множиною даних. Узагальнений алгоритм, що реалізує підхід на основі попередньої кластеризації, має наступний вигляд (рис. 2):

Рисунок 2 – Блок-схема алгоритму попередньої кластеризації даних

На першому кроці відбувається відбір цільових показників залежно від задач аналізу. Для спрощення обробки даних, що мають високу розмірність, використовується процедура факторного аналізу. До отриманої множини даних застосовується метод кластерного аналізу, після чого відбувається формування n множин даних, де n – кількість виявлених кластерів. На виході алгоритму отримують набори даних, що містять меншу кількість унікальних значень і придатні для подальшого аналізу.

Ефективність застосування даного підходу як етапу попередньої обробки даних будемо оцінювати у процесі відновлення пропущених значень з допомогою декількох алгоритмів імпутації. Розглянемо далі моделі, які реалізують процес відновлення пропущених даних.

Моделі імпутації даних. Як було зазначено вище, алгоритм обробки показників із акаунтів соціальних мереж повинен бути пристосованим до роботи з даними змішаної природи. Крім того, враховуючи наявність фізичних зв'язків між акаунтами і можливість існування очевидних або прихованих зв'язків між даними, прийнято рішення про застосування методів імпутації на модельній основі [3]. У роботі автори пропонують п'ять моделей імпутації на основі алгоритмів машинного навчання, два з яких дозволяють одночасну обробку змінних різної природи (асоціативні правила й випадкові ліси), два використовують номінальні атрибути лише в якості міток класів (машини опорних векторів і нейронна мережа), а останній не включає у процес аналізу нечислові дані (EM-алгоритм), однак є одним із найпопулярніших методів імпутації [7]. Опишемо детально кожну з моделей, передбачаючи застосування підходу на основі кластеризації на етапі попередньої обробки даних.

1. Модель на основі асоціативних правил (association rules), AR. Асоціативні правила – набір спеціальних правил, які дозволяють знаходити й описувати закономірності у великих наборах даних [8]. У випадку вирішення задачі імпутації показників із акаунтів соціальної мережі застосування методу пошуку асоціативних правил має наступні переваги: можливість одночасної обробки даних змішаної природи і врахування їх зв'язності в силу механізму пошуку асоціацій. Робота моделі основана на знаходженні очевидних і прихованих паттернів у вихідному наборі даних і їх подальшому використанні для відновлення пропущених значень. Алгоритм, що описує роботу моделі AR, має наступний вигляд (рис. 3):

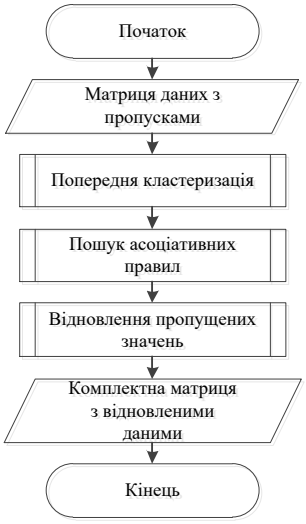

Рисунок 3 – Блок-схема алгоритму роботи моделі імпутації на основі асоціативних правил

На першому кроці реалізується запропонований підхід до попередньої обробки даних з допомогою процедури кластеризації. Далі у кожній із отриманих матриць виконується пошук асоціацій. На етапі відновлення пропущених значень відбувається пошук придатних для імпутації правил. Якщо для поточного пропуску таке правило існує, відбувається його заміна значенням зі слідуванням правила, інакше використовується значення з найбільшою частотою (most common value).

Принцип роботи наступних трьох моделей на основі випадкового лісу, машини опорних векторів і нейронної мережі полягає в навчанні на комплектних даних шляхом застосування одного з алгоритмів і подальшому використанні навченої моделі для відновлення пропущених значень.

2. Модель на основі випадкових лісів (random forests), RF. Випадковий ліс являє собою класифікатор, який складається з набору класифікаторів на основі дерев рішень $\{h(x,\Theta_k), k=1,\ldots\}$, де x – вхідний вектор даних, $h(x,\Theta_k)$ – дерева рішень, в якому кожне дерево голосує за один із класів входу x [9]. Кожне з дерев рішень будується за вибіркою, отриманою з вихідної множини даних з допомогою процедури бутстрепа. Фінальна класифікація виконується на основі голосування моделей.

Випадкові ліси мають ряд суттєвих переваг: захист від «перепідгонки» (overfitting) моделі, стійкість до шумів і викидів у даних, можливість обробки показників, що вимірюються в різних шкалах, – які обґрунтовують доцільність їх застосування у випадку обробки даних із соціальних мереж. Алгоритм, що описує роботу моделі RF, має наступний вигляд (рис. 4):

Рисунок 4 – Блок-схема алгоритму роботи моделі імпутації на основі випадкових лісів

Як і для попередньої моделі, на першому кроці відбувається реалізація підходу до обробки даних з допомогою процедури кластеризації. Далі вихідна матриця «об'єкт-ознака» розділяється на дві підмножини: навчальну, що містить лише комплектні дані, і тестову, з пропущеними значеннями. Перша підмножина є основою для застосування алгоритму випадкового лісу і навчання моделі. Відновлення пропущених значень виконує уже налаштована модель, приймаючи на вхід тестову підмножину з пропусками.

3. Модель на основі машин опорних векторів (support vector machines), SVM. Машина опорних векторів – алгоритм, що використовується для вирішення класифікаційних і регресійних задач, в

якому кожен об'єкт аналізованих даних представляється у вигляді вектору в p – вимірному просторі [10]. Метою методу є переміщення вихідних векторів у простір вищої розмірності і пошук гіперплощини, яка розділяє їх з максимальним проміжком. Класифікація об'єктів відбувається на основі їх відображення у той самий простір і віднесення до певних категорій відповідно до того, в яку частину гіперплощини вони потрапляють.

До переваг застосування алгоритму машин опорних векторів відносять ефективність у багатовимірному просторі даних і багатофункціональність внаслідок наявності ряду функцій ядра. Основним недоліком є можливість обробки лише числових даних і необхідність нормалізації вхідних показників. Однак, не зважаючи на недоліки, метод опорних векторів є перспективним для імпутації пропущених значень, оскільки може виявляти нелінійні функціональні зв'язки між даними. Алгоритм, що описує роботу моделі SVM, має наступний вигляд (рис. 5):

Рисунок 5 – Блок-схема алгоритму роботи моделі імпутації на основі машин опорних векторів

Як і у випадку попередньої моделі, після процедури кластеризації на основі вихідної матриці об'єкт-ознака відбувається формування навчальної і тестової підмножин. Оскільки алгоритм, який лежить в основі моделі, працює тільки з числовими даними, а номінальні здатен оброблювати у вигляді міток класів, на етапі формування підмножин відбувається виключення нечислових показників із навчальних наборів даних, якщо вони існують. На наступному етапі відбувається нормалізація даних множин, на яких здійснюється навчання моделі й відновлення на її основі пропущених значень.

4. Модель на основі штучної нейронної мережі (artificial neural network), ANN. Виходячи зі складної природи й наявності зв'язків між даними предметної області соціальних мереж, а також числової природи переважної більшості атрибутів, розглянемо модель імпутації на основі штучної нейронної мережі.

Нейронна мережа – паралельно розподілена система простих процесорних елементів (нейронів), які здатні виконувати найпростішу обробку даних і тісно пов'язані між собою зваженими зв'язками [11]. Особливості архітектури мережі та її функціонування визначаються моделлю нейронів, їх активаційною функцією, топологією, кількістю шарів і алгоритмом навчання. До переваг застосування нейронних мереж відносять: можливість навчання, здатність знаходження складних нелінійних закономірностей, стійкість до шумів і викидів, адаптивність навчання, масштабованість внаслідок паралельної структури. Серед недоліків варто зазначити складність налаштування численних початкових параметрів навчання й розподілу ваг і схильність до знаходження локальних мінімумів під час навчання.

Враховуючи широкий ряд параметрів для налаштування і можливих конфігурацій нейронних мереж, будемо розглядати їх застосування з точки зору обробки некомплектних даних. Базовою архітектурою при вирішенні подібних задач звичайно виступає багатошаровий перцептрон із одним або двома прихованими шарами й активаційною функцією гіперболічного тангенсу або сигмоїди [12, 13]. Головне обмеження, яке накладається на конфігурацію мережі – відповідність розмірності вектора вхідних даних кількості нейронів у вхідному шарі та розмірності вихідного вектора кількості вихідних нейронів. Алгоритм, що описує роботу моделі ANN, має наступний вигляд (рис. 6):

Рисунок 6 – Блок-схема алгоритму роботи моделі імпутації на основі нейронної мережі

Робота моделі ANN також починається з процедури кластеризації вихідної матриці даних і формування навчальної і тестової множин. Далі відбувається нормалізація даних, побудова мережі згідно з

визначеною конфігурацією і параметрами її навчання. На останньому етапі на вхід навченої мережі подається тестова множина даних і відбувається імпутація пропущених значень.

5. Модель на основі EM-алгоритму (EM-algorithm), EM. EM алгоритм – метод, який дозволяє знаходити оцінки максимальної правдоподібності в параметричних моделях для некомплектних даних [14]. Він являє собою ітеративну процедуру, яка включає два кроки E і M: обчислення математичного очікування (expectation) і максимізація (maximization). Звичайно на кроці E виконується обчислення очікуваного значення суми змінних з пропусками з припущенням, що відомі параметри середнього значення і коваріаційна матриця. На кроці M очікуване значення суми змінних використовується для отримання середнього значення і матриці коваріацій. На кожному кроці робиться припущення, що є значення одного з невідомих параметрів, і на його основі обчислюється інший. Подібне ітеративне повторення кроків E і M повторюється до збіжності процесу, тобто допоки параметри, які змінюються в процесі роботи алгоритму, не перестануть суттєво змінюватися [1]. Основним обмеженням застосування EM-алгоритму є припущення про багатовимірний нормальний розподіл аналізованих даних [15–16]. Однак, не зважаючи на це, EM-алгоритм є одним із найбільш поширених модельних методів імпутації. Для оцінки ефективності застосування даного методу для даних із акаунтів соціальних мереж розробимо модель імпутації на його основі. Алгоритм, що описує роботу моделі EM, має наступний вигляд (рис. 7):

Рисунок 7 – Блок-схема алгоритму роботи моделі імпутації на основі EM-алгоритму

Робота моделі EM починається з процедури кластеризації вихідної матриці даних і формування навчальної і тестової множин. Далі виконується обчислення параметрів пропущених значень на основі комплектних даних і заміна пропусків знайденими оцінками. Після цього відбувається максимально правдоподібне оцінювання знайдених параметрів і заповнення пропущених значень.

Результати застосування розроблених моделей. Підходи до оцінки похибок імпутації. До оцінки ефективності роботи розроблених моделей імпутації пропонуються два підходи. Перший – на основі оцінки частоти коректно відновлених значень, який є універсальним і може застосовуватись як для номінальних, так і для числових змінних. Другий метод оснований на обчисленні середньої квадратичної похибки і може застосовуватися тільки для числових змінних.

Нехай I_{imp} – вектор відновлених значень деякого атрибуту K, I_{real} – вектор реальних значень, n_I – довжина векторів. Частоту неправильно відновлених значень позначимо як n_{I_false}, частоту коректно відновлених як n_{I_true}, так що $n_{I\_false} + n_{I\_true} = n_I$. У випадку необхідності оцінки імпутації даних з допустимим відхиленням відновлених значень від абсолютних у деякому інтервалі Δr частота коректно відновлених даних обчислюватиметься як:

$$n_{I\_true} = \sum_{i=1}^{n_I} I_{imp_i} \in \left[I_{real_i} - \Delta r; I_{real_i} + \Delta r \right]. \qquad (1)$$

Для оцінки якості імпутації пропусків на основі частоти коректно відновлених значень застосуємо відносну характеристику:

$$\delta = \frac{n_{I\_true}}{n_I} \times 100\% . \qquad (2)$$

Оцінка якості відновлення пропущених даних на основі обчислення середньоквадратичної похибки також є одним із найбільш розповсюджених підходів [17–18] у сфері імпутації. Вона застосовується тільки для оцінки числових значень, має широкий ряд модифікацій, але найчастіше основана на обчисленні нормалізованого квадрату середньоквадратичної похибки й описується наступним виразом [19]:

$$NRMSE = \sqrt{\frac{mean\left(\left(I_{imp} - I_{real}\right)^2\right)}{\mathrm{var}\left(I_{real}\right)}}, \qquad (3)$$

де I_{imp} – вектор відновлених значень, I_{real} – вектор реальних значень, *mean* – вибіркове середнє значення вектору квадратів різниці між реальними і відновленими даними, *var* – дисперсія значень вектору реальних даних. Значення NRMSE, близькі до 0, свідчать про високу ефективність процесу імпутації і близькість вектору відновлених даних до вектору реальних, і навпаки, зі зростанням значення показника NRMSE знижується якість імпутації.

Дослідження ефективності застосування розроблених моделей. Під час дослідження роботи запропонованих моделей імпутації розглянемо ефективність застосування підходу на основі попередньої кластеризації. Для цього у процесі відновлення пропущених даних будемо застосовувати два види кожної з моделей: із етапом кластеризації, алгоритми роботи яких описані вище, та аналогічні, але без попередньої обробки.

Крім того, виконаємо перевірку гіпотези щодо впливу матриці збагачення на якість результатів імпутації. Для цього розроблені моделі будемо застосовувати на наборах даних двох видів: некомплектній і розширеній матрицях. Оскільки вихідні дані з акаунтів соціальної мережі містять пропуски, для коректної оцінки якості імпутації сформуємо на їх основі дві комплектні модельні множини даних шляхом видалення всіх пропущених значень. Далі випадковим чином згенеруємо в них різні відсотки штучних пропусків (1, 2, 5, 10, 20, 30, 50, 70 %) і застосуємо до отриманих некомплектних наборів розроблені моделі. Маючи множини відомих реальних значень, які були видалені, і результати моделей, виконаємо оцінку якості імпутації.

Таким чином, у процесі проведення експерименту будемо досліджувати два види кожної з моделей на двох множинах даних. На діаграмах, що представляють результати імпутації, введені наступні позначення: крива All_data Only miss – модель без кластеризації на некомплектній множині, Clusters Only miss – модель з етапом кластеризації на некомплектній множині, All_data Extended – модель без кластеризації на розширеній матриці даних, Clusters Extended – модель з етапом кластеризації на розширеній матриці даних. Розглянемо спочатку результати імпутації номінального атрибуту «Сімейний стан». На рис. 8–11 представлені криві похибки δ досліджуваних моделей імпутації на основі асоціативних правил, випадкового лісу, машини опорних векторів і нейронної мережі для різних відсотків пропущених значень:

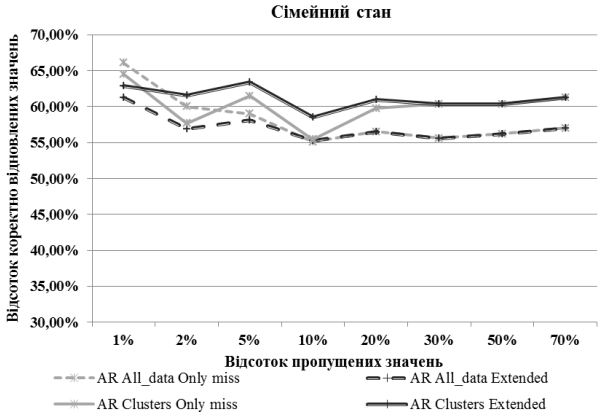

Рисунок 8 – Результати відновлення пропущених значень атрибуту V2 з допомогою моделі на основі асоціативних правил (AR)

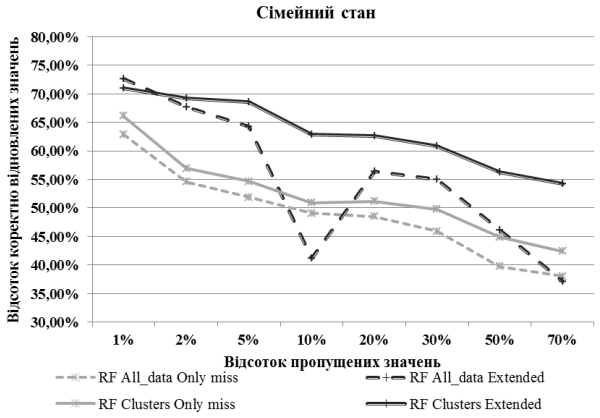

Рисунок 9 – Результати відновлення пропущених значень атрибуту V2 з допомогою моделі на основі випадкового лісу (RF)

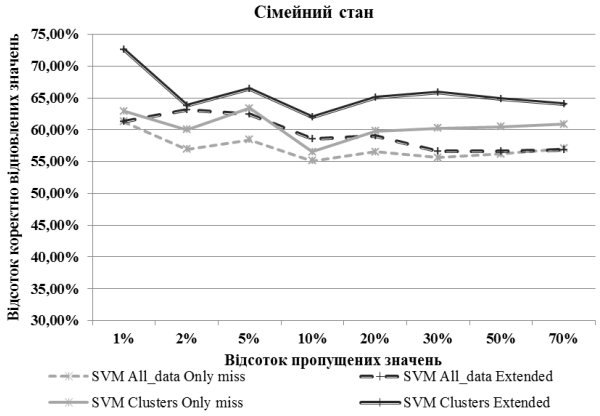

Рисунок 10 – Результати відновлення пропущених значень атрибуту V2 з допомогою моделі на основі машини опорних векторів (SVM)

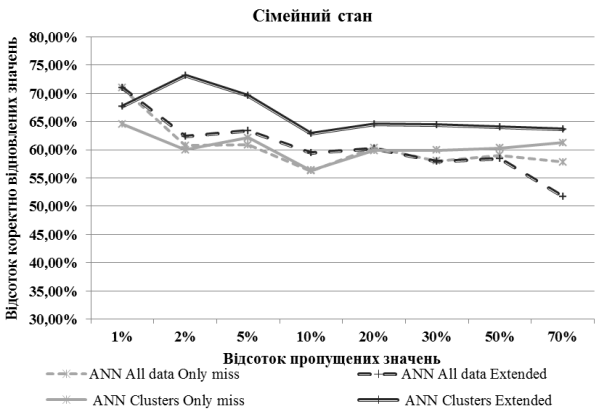

Рисунок 11 – Результати відновлення пропущених значень атрибуту V2 з допомогою моделі на основі нейронної мережі (ANN)

Розглянемо спочатку вплив процедури кластеризації на результати імпутації. Група пунктирних кривих описує моделі, що не мають етапу попередньої кластеризації, суцільні – із включенням даного етапу. Як видно з рисунків, введення процедури кластеризації дозволяє отримувати кращі результати як при роботі з некомплектними множинами (група сірих кривих), так і з розширеними матрицями даних (група чорних кривих), підвищуючи коректність імпутації до 8%. При цьому також спостерігається вплив розширеної матриці даних на результати відновлення пропущених значень: її використання порівняно з некомплектною матрицею дозволяє підвищити коректність імпутації на 5–15 %.

Таким чином, експериментально встановлено, що при імпутації пропущених значень номінального атрибуту включення етапу попередньої кластериза-

ції і використання розширеної матриці даних цілком виправдані, оскільки призводять до стабільного для всіх моделей суттєвого підвищення відсотку коректно відновлених значень.

Отже, у подальшому будемо розглядати моделі імпутації номінальних атрибутів на основі асоціативних правил, випадкового лісу, машини опорних векторів і нейронної мережі, які включають етап попередньої кластеризації і застосовуються на основі розширеної матриці даних.

Перейдемо до розгляду роботи моделей на числових даних. Результати коректності імпутації будемо оцінювати з допомогою похибок на основі частоти коректно відновлених значень δ і NRMSE. На рис. 12–16 наведено похибки δ:

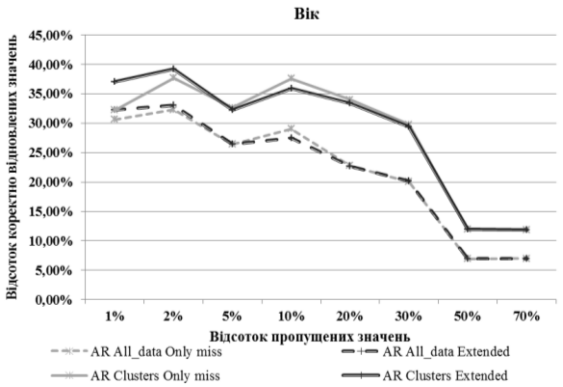

Рисунок 12 – Результати відновлення пропущених значень атрибуту V1 з допомогою моделі на основі асоціативних правил (AR)

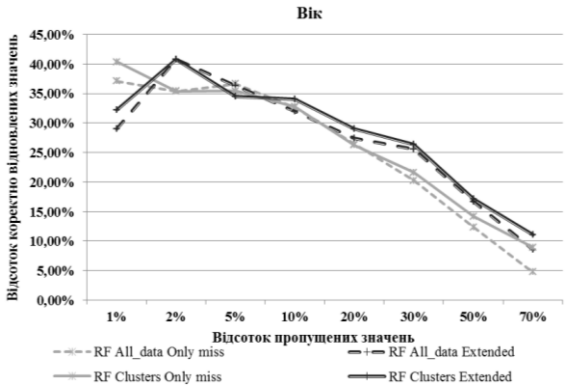

Рисунок 13 – Результати відновлення пропущених значень атрибуту V1 з допомогою моделі на основі випадкового лісу (RF)

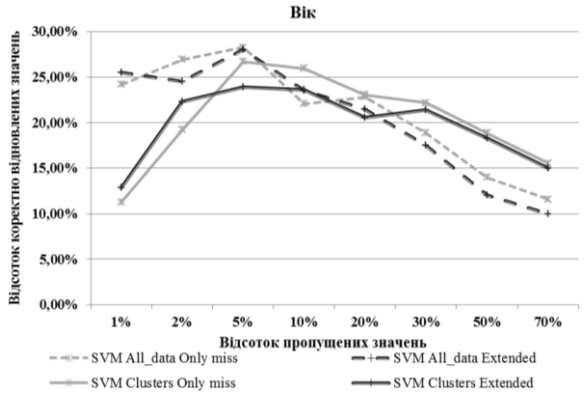

Рисунок 14 – Результати відновлення пропущених значень атрибуту V1 з допомогою моделі на основі машини опорних векторів (SVM)

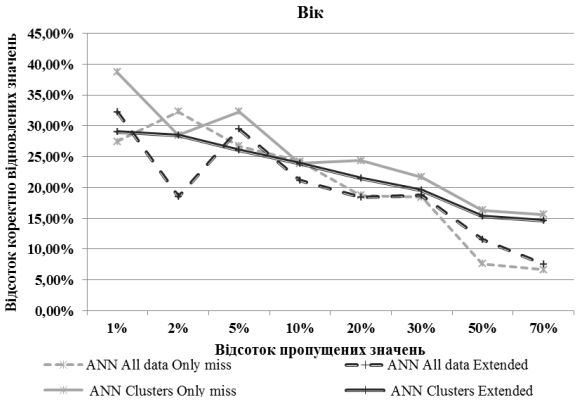

Рисунок 15 – Результати відновлення пропущених значень атрибуту V1 з допомогою моделі на основі нейронної мережі (ANN)

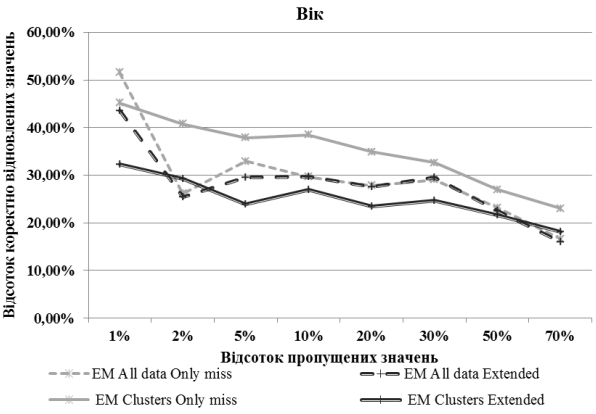

Рисунок 16 – Результати відновлення пропущених значень атрибуту V1 з допомогою моделі на основі ЕМ-алгоритму (EM)

Виконаємо порівняння роботи моделей з етапом попередньої кластеризації і без. Як видно із рисунків, введення процедури кластеризації у цілому дозволяє отримати кращі результати імпутації для моделей AR, ANN і EM – підвищення відсотку коректності відновлення до 10 %.

На модель RF даний етап з точки зору похибки δ суттєво не впливає, а на SVM має негативний вплив при невисокому й середньому відсотках пропусків (від 1 до 20 %).

При дослідженні впливу розширеної матриці даних на ефективність імпутації виявлено, що для моделей AR і RF її використання не призводить до погіршення результатів, а у випадку роботи з високими частками пропусків дозволяє отримати вищий відсоток коректно відновлених даних для моделі RF (близько 5 %). Із результатів похибок моделей SVM, ANN і EM видно, що використання розширеної

матриці атрибутів спричиняє зниження якості імпутації, що може бути наслідком їх перепідгонки (overfitting). Таким чином, результати аналізу похибки δ доводять доцільність застосування етапу попередньої кластеризації для моделей AR, ANN і EM і використання розширеної матриці даних для моделей AR і RF. Для більш об'єктивної оцінки ефективності роботи моделей розглянемо значення похибки імпутації NRMSE (рис. 17–21).

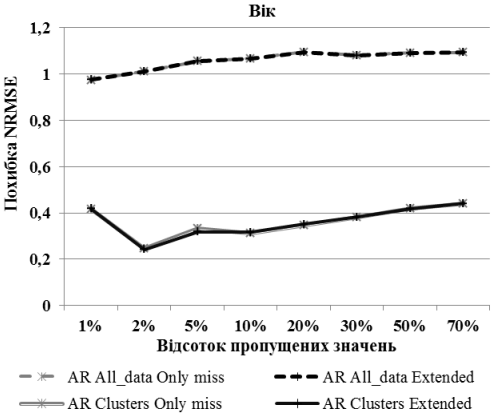

Рисунок 17 – Похибка відновлення NRMSE значень атрибуту V1 з допомогою моделі на основі асоціативних правил (AR)

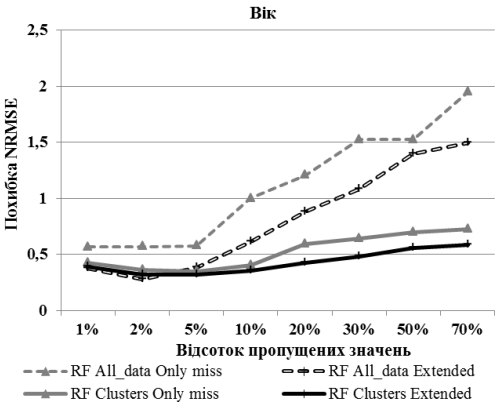

Рисунок 18 – Похибка відновлення NRMSE значень атрибуту V1 з допомогою моделі на основі випадкового лісу (RF)

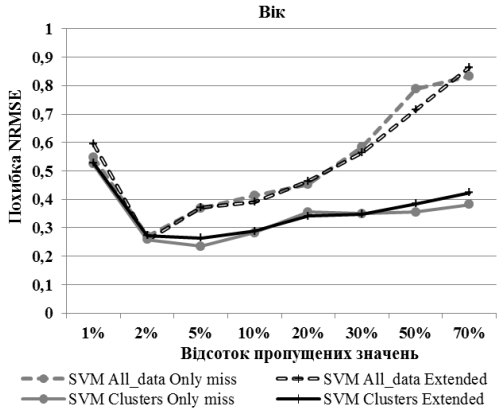

Рисунок 19 – Похибка відновлення NRMSE значень атрибуту V1 з допомогою моделі на основі машини опорних векторів (SVM)

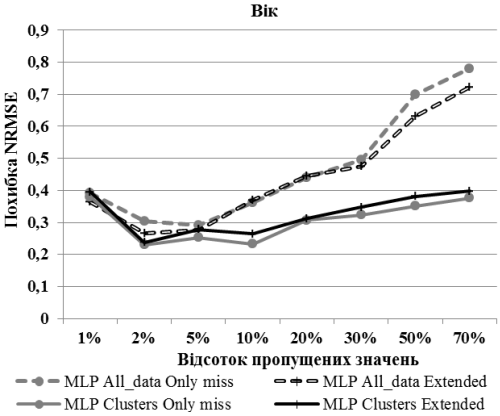

Рисунок 20 – Похибка відновлення NRMSE значень атрибуту V1 з допомогою моделі на основі нейронної мережі (ANN)

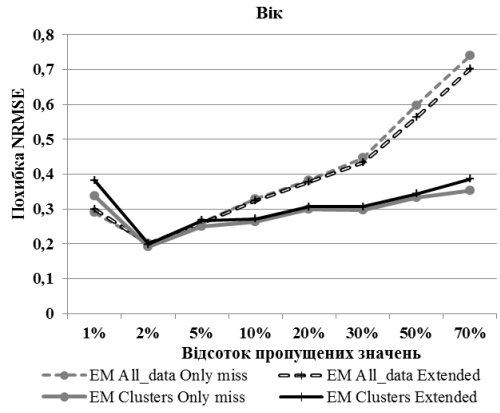

Рисунок 21 – Похибка відновлення NRMSE значень атрибуту V1 з допомогою моделі на основі ЕМ-алгоритму (EM)

Як видно із наведених рисунків, при оцінці впливу розширеної матриці на ефективність імпутації значення похибки NRMSE узгоджуються з оцінками δ і підтверджують, що доповнення некомплектної матриці атрибутами матриці збагачення не призводить до зростання або дещо зменшує значення похибки NRMSE для моделей AR і RF, при цьому практично не впливаючи на похибки моделей SVM, ANN і EM. Враховуючи отримані значення двох похибок, розширену матрицю даних доцільно використовувати лише при роботі з моделлю RF, оскільки вона не впливає на модель AR і може призводити до погіршення результатів роботи моделей SVM, ANN і EM.

Крім того, отримані результати доводять доцільність введення етапу попередньої кластеризації вихідної множини даних для всіх моделей імпутації числових атрибутів, оскільки застосування такого підходу дозволяє уникнути різкого зростання зна-

чень похибки NRMSE зі збільшенням відсотку пропущених значень. Приймаючи до уваги значення обох похибок імпутації, обґрунтованим є вибір наступних моделей відновлення числових даних для подальшого застосування: моделі на основі випадкового лісу з етапом попередньої кластеризації і розширеною матрицею даних і моделей на основі асоціативних правил, машини опорних векторів, нейронної мережі й EM-алгоритму з етапом попередньої кластеризації, що навчаються на некомплектних матрицях.

ВИСНОВКИ. Виконано дослідження структури апріорно-апостеріорних даних, які містять акаунти користувачів соціальної мережі, й обґрунтовано відбір цільових показників для подальшого аналізу. Показано, що множину показників, що описують користувача, можна розділити на 2 групи: атрибути, які можуть містити пропущені дані (некомплектна матриця), і зажди комплектні атрибути (матриця збагачення), і на їх основі сформульовано поняття розширеної матриці даних. Висунуто гіпотезу щодо впливу матриці збагачення на якість результатів імпутації.

Для вирішення проблеми значної кількості унікальних значень розроблено підхід на основі попередньої кластеризації вихідної множини даних. Для перевірки гіпотези щодо впливу матриці збагачення на ефективність відновлення пропущених значень запропоновано два варіанти формування множин даних для навчання моделей імпутації: на основі суто некомплектної матриці і з використанням розширеної матриці даних.

Виходячи з характерних особливостей даних із акаунтів соціальної мережі розроблено п’ять моделей імпутації на основі алгоритмів машинного навчання з використанням підходів на основі попередньої кластеризації і розширеної матриці даних. Запропоновано два підходи до оцінки якості роботи моделей відновлення пропусків на основі обчислення частоти коректно відновлених значень δ і середньої квадратичної похибки NRMSE. Виконано дослідження ефективності застосування розроблених моделей для номінальних і числових даних.

Доведено позитивний вплив етапу попередньої кластеризації при обробці номінальних і числових даних, який дозволяє підвищувати коректність відновлення пропусків до 10 % і знижувати похибку NRMSE на 0,1–0,7.

Виявлено, що використання розширеної матриці даних має різний ефект при застосуванні моделей для номінальних і числових даних. Встановлено, що навчання моделей на некомплектній матриці, доповненій матрицею збагачення, дозволяє підвищити коректність імпутації номінальних значень на 5–15 %. При цьому під час роботи моделей на числових даних використання розширеної матриці даних позитивно впливає лише на коректність імпутації моделі RF (підвищення δ до 5 % і зменшення NRMSE на 0,1–0,2), при цьому ніяк не відображаючись на моделі AR і призводячи до погіршення роботи моделей SVM, ANN і EM – зниження коректності відновлення у середньому на 5–10 %.

Обґрунтовано доцільність використання декількох модифікацій моделей для імпутації атрибутів різної природи. Встановлено, що найбільш ефективними при відновленні номінальних атрибутів є моделі на основі асоціативних правил (AR), випадкового лісу (RF), машини опорних векторів (SVM) і нейронної мережі (ANN), які включають етап попередньої кластеризації і застосовуються на основі розширеної матриці даних. При імпутації числових даних найоптимальнішими є модель на основі випадкового лісу з етапом попередньої кластеризації і розширеною матрицею даних і моделі на основі асоціативних правил, машини опорних векторів, нейронної мережі й EM-алгоритму з етапом попередньої кластеризації, що навчаються на некомплектних матрицях. Враховуючи неоднорідність значень похибок для розроблених моделей, інтерес для подальшого дослідження представляє розробка методу імпутації, який зможе виконувати відновлення пропущених даних більш стабільно і якісно. Вирішення даної задачі можливе шляхом синтезу ансамблів моделей імпутації, кандидатами на включення до яких є розроблені моделі.

ЛІТЕРАТУРА

1. Graham J. W. Missing Data: Analysis and Design. – New York: Springer, 2012. – 324 p.

2. Znidarsic A., Ferligoj A., Doreian P. Non-response in social networks: The impact of different non-response treatments on the stability of blockmodels // Social Networks. – 2012. – Iss. 34. – PP. 438–450.

3. Литтл Р.А., Рубин Д.Б. Статистический анализ данных с пропусками / Пер. с англ. – М: Финансы и статистика, 1990. – 336 с.

4. McKnight P., McKnight K., Sidani S., Figueredo A. Missing Data: A Gentle Introduction. – New York: Guilford Press, 2007. – 250 p.

5. Слабченко О.О., Сидоренко В.Н. Улучшение качества исходных данных в задачах моделирования интернет-сообществ на основе комплексного применения моделей сегментации, импутации и обогащения данных // Вісник Кременчуцького національного університету імені Михайла Остроградського. – 2013. – Вип. 6/2013 (83). – С. 50–58.

6. Slabchenko O., Sydorenko V., Siebert X. An improved algorithm for imputation data from social network accounts with use of association rules // Збірник матеріалів XXI Міжнародної науково-технічної конференції студентів, аспірантів та молодих учених КрНУ імені Михайла Остроградського «Актуальні проблеми життєдіяльності суспільства», 24–25 квітня, м. Кременчук. – Україна, 2014. – С. 45–46.

7. Dempster A., Rubin D. Maximum likelihood from incomplete data via the EM algorithm // Journal of the Royal Statistical Society. – 1977. – Vol. 39(1). – PP. 1–38.

8. Agrawal R., Srikant R. Fast Algorithms for Mining Association Rules in Large Databases // Proceedings of the 20th International Conference on Very Large Data Bases. – 1994. – PP. 487–499.

9. Breiman L. Random forests // Machine Learning. – 2001. – Vol. 45(1). – PP. 5–32.

10. Cortes C., Vapnik V. Support-Vector Networks // Machine Learning. – 1995. – Vol. 20(3). – PP. 273–297.

11. Hiregoudar S., Manjunath K., Patil K. A survey: research summary on neural networks // International Journal of Research in Engineering and Technology. – 2014. – Vol. 3(3). – PP. 385–389.

12. Silva-Ramírez E., Pino-Mejías R., López-Coello M. Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns // Applied Soft Computing. – 2015. – Vol. 29. – PP. 65–74.

13. Nelwamondo F., Mohamed S., Marwala T. Missing data: A comparison of neural network and expectation maximization techniques // Current Science. – 2007. – Vol. 93(11). – PP. 1514–1521.

14. Enders C.K. A Primer on the Use of Modern Missing-Data Methods in Psychosomatic Medicine Research // Psychosomatic Medicine. – 2006. – Vol. 68(3). – PP. 427–436.

15. Schafer J. L. Analysis of Incomplete Multivariate Data. – New York: Chapman and Hall/CRC, 1997. – 444 p.

16. Little J. A., Rubin D. B. Statistical Analysis with Missing Data, 2nd Edition. – New Jersey: John Wiley & Sons, 2002. – 408 p.

17. Aydilek I. B., Arslan A. A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm // Information Sciences. – 2013. – Vol. 233. – PP. 25–35.

18. Ferrari P. A., Annoni P., Barbiero A., Manzi G. An imputation method for categorical variables with application to nonlinear principal component analysis // Computational Statistics & Data Analysis. – 2011. – Vol. 55(7). – PP. 2410–2420.

19. Oba S., Sato M., Takemasa I., Monden M., Matsubara K., Ishii, S.A Bayesian missing value estimation method for gene expression profile data // Bioinformatics. – 2003. – Vol. 19(16). – PP. 2088–2096.

THE EFFECTIVENESS OF METHODS AND ALGORITHMS FOR MISSING DATA IMPUTATION IN TASKS OF SOCIAL-NETWORK ANALYSIS

O. Slabchenko, V. Sydorenko

Kremenchuk Mykhailo Ostrohradskyi National University

vul. Pershotravneva, 20, Kremenchuk, 39600, Ukraine. E-mail: slabchenko.olesia@gmail.com

Purpose. Quality improvement of initial data from social networks users' accounts when preprocessing by application of methods for missing values imputation. **Methodology**. We have applied the data mining methods and algorithms (such as association rules, random forest, support vector machine, neural network and EM-algorithm) for development of models for imputation data from social networks users' accounts. We have employed methods of mathematical statistics to assess two errors of imputation for the proposed models. **Results**. We have researched the structure of a priori-a posteriori data from social networks users' accounts and justified a range of target attributes for further analysis. We have shown that the range of descriptive indicators can be divided into two groups: the ones that may contain missing data (incomplete matrix) and always complete attributes (enrichment matrix). We have formulated the concept of extended matrix of attributes and put forward a hypothesis concerning its effect on the quality of imputation. To solve the problem of the significant number of unique values in attributes we have proposed an approach on the basis of pre-clustering of the initial data. We have designed five models of imputation based on machine learning algorithms with application of pre-clustering method and extended matrix of attributes. We have proposed two methods to evaluate the quality of models on the basis of correctly imputed values rate δ and normalized root mean squared error (NRMSE). We have researched the efficiency of the proposed models for nominal and numerical data. Our experiments have proven positive effect of pre-clustering at processing of different data types. They have also found out that usage of extended matrix impacts on imputation process: it enhances efficiency of models in case of nominal data treatment and can result in worsening in case of numerical data. We have justified and selected the optimal versions of models for imputation of different data types for further application. **Originality**. We have developed a method of reducing the number of unique values in attributes on basis of pre-clustering that enables to increase correctness of missing values imputation up to 10% and decrease NRMSE error to 0.1–0.7. We have designed the approach to imputation on basis of extended matrix, which, unlike existing ones, enables to supplement incomplete matrix with matrix of always complete indicators and thus to increase the informativity of initial data at modelling. It was shown that the use of extended matrix enables to increase correctness of nominal data imputation by 5–15 %. **Practical value**. When developing models of imputation, we have proposed the approach to analysis of data from social networks users' accounts on the basis of extended matrix. This makes possible to take into account always complete attributes in order to increase effectiveness of missing values of imputation process. We have designed the method of reducing the number of unique values in attributes, which enables to train imputation models more effectively. The introduced models can be used for the construction of ensembles of models in order to improve stability and quality of imputation. References 19, figures 21, tables 1.

Key words: imputation, extended matrix, pre-clustering, association rules, random forest, support vector machine, neural network, EM-algorithm.

REFERENCES

1. Graham, J. W. (2012), *Missing Data: Analysis and Design*, Springer, New York, USA.

2. Znidarsic, A., Ferligoj, A., Doreian, P. (2012), "Non-response in social networks: The impact of different non-response treatments on the stability of blockmodels", *Social Networks*, Iss. 34, pp. 438–450.

3. Little, A., Rubin, D. (1990), *Statistical analysis with missing data* [Statisticheskiy analiz dannyih s propuskami], Finansy i statistica, Moscow, Russia.

4. McKnight, P., McKnight, K., Sidani, S., Figueredo, A. (2007), *Missing Data: A Gentle Introduction*, Guilford Pres, New York, USA.

5. Slabchenko, O., Sydorenko, V. (2013), "The improvement of initial data quality in modeling ploblems of online communities on the base of combined implementation of segmentation, imputation and data enrichment models", *Transactions of Kremenchuk Mykhailo Ostrohradskyi National University*, Vol. 6, iss. 83, pp. 50–58.

6. Slabchenko, O., Sydorenko, V., Siebert, X. (2014), "An improved algorithm for imputation data from social network accounts with use of association rules", *Materialy XXI Mezhdunarodnoy nauchno-prakticheskoy konferentsii studentov, aspirantov i molodykh uchenykh KrNU imeni Mikhaila Ostrogradskogo "Aktual'nyye problemy zhiznedeyatel'nosti obshchestva"* [Proceedings of the 21 International scientific and practical conference of students, PhD students and young scientists of Kremenchuk Mykhailo Ostrohradskiy National University on Actual problems of society activity], April 24–25, Kremenchug, pp. 45–46.

7. Dempster, A., Rubin, D. (1977), "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, vol. 39(1), pp. 1–38.

8. Agrawal, R., Srikant, R. (1994), "Fast Algorithms for Mining Association Rules in Large Databases", *Proceedings of the 20th International Conference on Very Large Data Bases*, pp. 487–499.

9. Breiman, L. (2001), "Random forests", *Machine Learning*, Vol. 45(1), pp. 5–32.

10. Cortes, C., Vapnik, V. (1995), "Support-Vector Networks", *Machine Learning*, Vol. 20(3), pp. 273–297.

11. Hiregoudar, S. B., Manjunath, K., Patil, K. S. (2014), "A survey: research summary on neural networks", *International Journal of Research in Engineering and Technology*, Vol. 3(3), pp. 385–389.

12. Silva-Ramírez, E. L., Pino-Mejías, R., López-Coello, M. (2015), "Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns", *Applied Soft Computing*, Vol. 29, pp. 65–74.

13. Nelwamondo, F. V., Mohamed, S., Marwala, T. (2007), "Missing data: A comparison of neural network and expectation maximization techniques", *Current Science*, Vol. 93(11), pp. 1514–1521.

14. Enders, C. K. (2006), "A Primer on the Use of Modern Missing-Data Methods in Psychosomatic Medicine Research", *Psychosomatic Medicine*, Vol. 68(3), pp. 427–436.

15. Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, Chapman and Hall, New York, USA.

16. Little, J. A., Rubin, D. B. (2002), *Statistical Analysis with Missing Data, 2nd Edition*, John Wiley & Sons, New Jersey, USA.

17. Aydilek, I. B., Arslan, A. (2013), "A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm", *Information Sciences*, Vol. 233, pp. 25–35.

18. Ferrari, P. A., Annoni, P., Barbiero, A., Manzi, G. (2011), "An imputation method for categorical variables with application to nonlinear principal component analysis", *Computational Statistics & Data Analysis*, Vol. 55(7), pp. 2410–2420.

19. Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K., Ishii, S. (2003), "A Bayesian missing value estimation method for gene expression profile data", *Bioinformatics*, Vol. 19(16), pp. 2088–2096.

Стаття надійшла 11.04.2016.