Вісник Національного технічного університету "ХПІ", 2018, № 42 (1318)
ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

УДК 004.832.34+303.442.3 DOI: 10.20998/2411-0558.2018.42.02

Т. О. БІЛОБОРОДОВА, канд. техн. наук, ст. викл., СНУ ім. В. Даля, Сєвєродонецьк
І. С. СКАРГА-БАНДУРОВА, д-р техн. наук, доц., зав.каф., СНУ ім. В. Даля, Сєвєродонецьк

КОМПЛЕКСНИЙ ПІДХІД ДО ОБРОБКИ РІЗНОРІДНИХ МЕДИЧНИХ ДАНИХ З ВІДСУТНІМИ ЗНАЧЕННЯМИ

Показана необхідність врахування змішаних наборів даних з відсутніми значеннями. Запропоновано узагальнений підхід до обробки різнорідних медичних даних з урахуванням типу даних, механізму їх відсутності. Проведене опрацювання трьох різних наборів даних з використанням запропонованого підходу та контрольної моделі. Надано порівняння ефективності опрацювання даних. Іл.: 2. Табл.: 2. Бібліогр.: 12 назв.

Ключові слова: різнорідні медичні дані, відсутні значення, обробка, контрольна модель.

Опис проблеми та аналіз літератури. За даними звіту про тенденції в галузі охорони здоров'я від Stanford Medicine [1] майбутнє охорони здоров'я залежить від ряду важливих тенденцій, серед яких виділяють прогнозування та профілактику захворюваності. У цьому контексті, якісний аналіз даних може потенційно поліпшити лікування пацієнтів, знайти невідомі фактори ризику захворювань або виявити супутні захворювання, зробити медичну діагностику більш точною, покращити управління витратами, тощо. Разом з тим, медичні дані є одними з найбільш складних в дослідженні типів даних [2].

В науковій літературі обговорення проблем медичних даних в основному фокусується на наявності помилок вимірювання, відсутніх значень, введення некоректних даних й т.і. Найбільш повний перелік основних недоліків медичних даних надано у [3]. Автори [4] розрізняють два типи проблем якості даних: неповні дані (відсутні та зсунені) і некоректні дані. В дослідженні [5] проблемами, що ускладнюють отримання якісних прогнозів станів пацієнтів реанімаційного відділення, визначені висока розмірність даних, їх незбалансованість та часова асинхронізація. В статті [6] також досліджуються дані пацієнтів реанімаційного відділення та основною проблемою якісного прогнозування майбутніх станів визначені відсутні дані. Проведений авторами [7] огляд провідних медичних журналів показав, що відсутні дані є звичайним явищем в рандомізованих дослідженнях з результатами, отриманими від пацієнтів. Відсутні значення також є загальною

проблемою в дослідженнях з поздовжніми, просторовими, багаторівневими або багатовимірними даними [8]. Іншою складністю медичних даних є їх різнорідність. Різнорідні медичні дані характеризуються неоднорідністю, незбалансованістю та зсувом значень в той чи інший бік відносно досліджуваної патології, що, власне і викликає труднощі для предиктивної аналітики. Для неоднорідних даних з відсутніми елементами можна практикувати різні підходи, але постійно є ризик застосування неефективної методики, що значно ускладнює отримання вагомих результатів [9]. Очевидно, що моделі, отримані при використанні даних незадовільної якості, будуть функціонувати неналежним чином. Таким чином, актуальною задачею є пошук підходу, який дозволить обирати найкращу модель для обробки різнорідних медичних даних, даних з відсутніми значеннями.

Метою роботи є розробка універсального підходу до обробки різнорідних даних з відсутніми значеннями та оцінка його ефективності.

Основна частина. Пропонований підхід до обробки різнорідних даних з відсутніми значеннями дозволяє врахувати механізми відсутності, тип і метод обробки і містить п'ять основних етапів (рис. 1).

На першому етапі проводиться якісна оцінка відсутніх даних. Визначаються типи змінних та механізм появи відсутніх значень. Другий етап містить кількісну оцінку відсутніх даних за результатами якої приймається рішення щодо відновлення даних або повного видалення наборів даних з відсутніми значеннями. На третьому етапі обирається метод обробки відсутніх даних. Детальний опис можливих стратегій надано у [10]. Четвертий етап містить перевірки зменшення набору даних після видалення спостережень або процедур оцінки чутливості після закінчення відновлення. На останньому етапі проводиться оцінка ефективності використаного методу обробки відсутніх даних з використанням формалізованих параметрів точності і мір ефективності відновлення відсутніх даних.

Набори даних. Для дослідження переваг запропонованого підходу до опрацювання відсутніх даних були використані три різних набори даних.

Перший набір містить 6 вхідних змінних (дані перебігу вагітності) та вихідну змінну (стан новонародженого). Набір містить 186 спостережень, з яких 81 з діагнозом "патологія".

Другий набір даних містить клінічні дані та дані мамограм. Набір містить 5 вхідних змінних: 1 кількісну та 4 порядкові, та вихідну змінну – відсутність або наявність злоякісного новоутворення молочної залози.

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

Рис. 1. Запропонована стратегія обробки різнорідних наборів даних з відсутніми значеннями

Таблиця 1

Інформація про досліджувані набори даних

Набір даних	Кількість спостере-жень	Кількість змінних	Типи змінних	Відсутні дані
Дані перебігу вагітності 12-38 тижнів вагітності	186	6	числові, порядкові, бінарні	так
Клінічні дані та результати мамограми пацієнток [11]	961	5	числові, порядкові, бінарні	так
Дані діагностики зображень одиночної протонної комп'ютерної томографії серця [12]	267	22	бінарні	ні

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

Третій набір містить дані діагностики зображень одиночної протонної комп'ютерної томографії серця (SPECT). Вихідна змінна поділяється на два класи: нормальні зображення і зображення з патологією. Дані складаються з оброблених зображень SPECT (пацієнтів). Дані були додатково оброблені для отримання 22 бінарних ознак зображення – вхідних змінних.

Процедура порівняння складалася з наступних етапів (рис. 2): перевірка набору даних на відсутні дані та, за необхідності, введення до 10% відсутніх значень для проведення дослідження, визначення механізму відсутності; визначення методу опрацювання відсутніх даних за запропонованою технологією та з використанням контрольної моделі; класифікація даних; оцінка результатів; порівняльний аналіз отриманих результатів.

Рис. 2. Процедура порівняння підходів до опрацювання різнорідних даних з відсутніми значеннями

В обраних наборах найбільший відсоток відсутніх даних спостерігається в наборі 1 (до 40%). Середній відсоток – в наборі 2 (понад 10%). Набір 3 не містив відсутні дані, отже, для експерименту, до нього штучно введено до 10% відсутніх значень. В якості контрольної моделі для обробки відсутніх значень використано модель відновлення медіаною прилеглих значень. Змінні, відновлення яких не відповідають відновленню медіаною прилеглих значень, залишені відсутніми.

В результаті використання контрольної та запропонованої в роботі моделей отримані два повних набори даних. Кожен з наборів випадковим чином розділено на навчальний набір (82%) та тестовий набір (18%). Класифікацію проведено з використанням алгоритму Random Tree. На підставі кількості помилок класифікації першого та другого роду, кількості істинно позитивних та істинно негативних спостережень розраховані коефіцієнт помилок класифікації та критерії ефективності класифікації для контрольної та розробленої моделі: чутливість, специфічність, точність.

Результати розрахунків критеріїв та коефіцієнту помилок класифікації для трьох досліджуваних наборів даних наведені в табл. 2. Найбільша розбіжність в параметрах якості отримана при відновленні набору даних 1, що містив велику кількість відсутніх значень.

Таблиця 2

Метрики якості для трьох наборів даних

Модель$^{\text{набір}}$	CER (%)	Чутливість (%)	Специфічність (%)	Точність (%)
Контрольна^1	45	62.5	52	54.54
Запропонована^1	27	70	76.9	72.7
Контрольна^2	32	73.19	63.88	68.02
Запропонована^2	27	76.76	67.12	72.67
Контрольна^3	27	87.8	23	72.1
Запропонована^3	25	87.8	30	72.2

За результатами розрахунків, запропонований підхід дає достатньо високі значення показників чутливості, специфічності, точності та мінімальний коефіцієнт помилок класифікації, що є дуже цінним при відновленні відсутніх значень, та істотно покращує властивості даних, що використовуються в подальшому аналізі.

Висновки. Представлено узагальнений підхід до обробки різнорідних даних з відсутніми значеннями, який враховує множину типів відсутніх значень, множину механізмів відсутності даних та множину методів опрацювання даних з відсутніми значеннями.

Проведена оцінка ефективності показала, шо даний підхід дає суттєве покращення показників чутливості, специфічності, точності та мінімальний коефіцієнт помилок класифікації. Варто відзначити, що для доведення ефективності пропонованої технології потрібно провести ще декілька серій експериментів з різними наборами даних та різними контрольними моделями. Подальші дослідження мають бути направлені на більш детальний підбір моделей опрацювання даних і організацію зворотного зв'язку між реальними, отриманими з інших клінічних джерел, і змодельованими результатами.

Список літератури:

1. School of Medicine: Stanford Medicine 2017 Health Trends Report Harnessing the Power of Data in Health, 2017. Режим доступу: https://med.stanford.edu/content/dam/sm/sm-news/documents/StanfordMedicine HealthTrendsWhitePaper2017.pdf (дата звернення 30.08.2018).

2. *Аврунін О.Г.* Сучасні інтелектуальні технології функціональної медичної діагностики: монографія / *О.Г. Аврунін, Є.В. Бодянський, М.В. Калашник, В.В. Семенець, В.О. Філатов.* – Харків: ХНУРЕ, 2018. – 236 с.

3. *Esfandiari N.* Knowledge discovery in medicine: Current issue and future trend / *N. Esfandiari, M.R. Babavalian, A.M.E. Moghadam, V.K. Tabar* // Expert Systems with Applications. – 2014. – Vol. 41 (9). – P. 4434-4463.

4. *Wu X.* Top 10 algorithms in data mining / *X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, S.Y. Philip, Z.H. Zhou* // Knowledge and information systems. – 2008. – Vol. 14 (1). – P. 1-37.

5. *Liu J.* Mortality prediction based on imbalanced high-dimensional ICU big data / *J. Liu, X.X. Chen, L. Fang, J.X. Li, T. Yang, Q. Zhan, K. Tong, Z. Fang* // Computers in Industry. – 2018. – Vol. 98. – P. 218-225.

6. *Nagrebetsky A.* Missing Data and ICU Mortality Prediction: Gone But Not to Be Forgotten. / *A. Nagrebetsky, E.A. Bittner* // Critical care medicine. – 2017. – Vol. 45 (12). – P. 2108-2109**.**

7. *Scharfstein D.* Final Report: Sensitivity Analysis Tools for Randomized Trials with Missing Data, 2017. – 112 p.

8. *Ringham B.M.* On the distribution of summary statistics for missing data / *B.M. Ringham, S.M. Kreidler, K.E. Muller and D.H. Glueck* // Communications in Statistics-Theory and Methods, 2018. – P. 1-17.

9. *Magnani M.* Techniques for dealing with missing data in knowledge discovery tasks. [Електронний ресурс] / *Magnani M.* // Режим доступу www URL: http://magnanim. web. cs. unibo. it/index. html (дата звернення 18.09.2018).

10. *Skarga-Bandurova I.* Strategy to Managing Mixed Datasets with Missing Items / *I. Skarga-Bandurova, T Biloborodova., Y. Dyachenko* // International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. Springer, Cham. – 2018. – P. 608-620.

11. *Elter M.* The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process / *M. Elter, R. Schulz-Wendtland, T. Wittenberg* // Medical Physics. – 2007. – Vol. 34 (11). – P. 4164-4172.

12. UCI Machine Learning Repository. [Електронний ресурс]. – Режим доступу: www URL: https://archive.ics.uci.edu/ml/datasets/SPECT+Heart (дата звернення 30.07.2018).

References:
1. School of Medicine: Stanford Medicine 2017 Health Trends Report Harnessing the Power of Data in Health (2017), available at: https://med.stanford.edu/ content/dam/sm/sm-news/documents/StanfordMedicine HealthTrendsWhitePaper2017.pdf (accessed 30 Aug 2018).
2. Avrunin, O.H., Bodianskyi, Ye.V., Kalashnyk, M.V., Semenets, V.V. and Filatov, V.O. (2018), *Suchasni intelektualni tekhnolohii funktsionalnoi medychnoi diahnostyky*, KhNURE, Kharkiv, 236 p.
3. Esfandiari, N., Babavalian, M.R., Moghadam, A.M.E. and Tabar, V.K. (2014), "Knowledge discovery in medicine: Current issue and future trend", *Expert Systems with Applications*, Vol. 41 (9), pp. 4434-4463.
4. Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Philip, S.Y. and Zhou, Z.H. (2008), "Top 10 algorithms in data mining", *Knowledge and information systems*, Vol. 14 (1), pp. 1-37.
5. Liu, J., Chen, X.X., Fang, L., Li, J.X., Yang, T., Zhan, Q., Tong, K. and Fang, Z. (2018), "Mortality prediction based on imbalanced high-dimensional ICU big data", *Computers in Industry*, Vol. 98, pp. 218-225.
6. Nagrebetsky, A. and Bittner, E.A. (2017), "Missing Data and ICU Mortality Prediction: Gone But Not to Be Forgotten", *Critical care medicine*, Vol. 45 (12), pp. 2108-2109**.**
7. Scharfstein, D. (2017), *Sensitivity Analysis Tools for Randomized Trials with Missing Data*, Final Report, 112 p.
8. Ringham, B.M., Kreidler, S.M., Muller, K.E. and Glueck, D.H. (2018), "On the distribution of summary statistics for missing data", *Communications in Statistics-Theory and Methods*, pp. 1-17.
9. Magnani, M. (2004), "Techniques for dealing with missing data in knowledge discovery tasks", available at: http://magnanim. web. cs. unibo. it/index. html, (accessed 18 Sept 2018).
10. Skarga-Bandurova, I., Biloborodova, T., and Dyachenko, Y. (2018), "Strategy to Managing Mixed Datasets with Missing Items", *Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Springer, pp. 608-620.
11. Elter, M., Schulz-Wendtland, R. and Wittenberg, T. (2007), "The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process", *Medical Physics*, Vol. 34 (11), pp. 4164-4172.
12. UCI Machine Learning Repository, available at: https://archive.ics.uci.edu/ml/ datasets/SPECT+Heart (accessed 30.07. 2018).

Статтю представив д.т.н., проф. Національного технічного університету "Харківський політехнічний інститут" С.Ю. Леонов

Надійшла (received) 13.08.2018

Biloborodova Tetyana, senior lecturer
Volodymyr Dahl East Ukrainian National University
59-a Central avenue, Severodonetsk, Luhansk region, Ukraine, 93400
Tel: (064) 522-89-97, e-mail: beloborodova.t@gmail.com
ORCID ID: 0000-0001-7561-7484

Skarga-Bandurova Inna, Dr. Sci. Tech., Assoc. Professor, Professor
Volodymyr Dahl East Ukrainian National University
59-a Central avenue, Severodonetsk, Luhansk region, Ukraine, 93400
Tel: (064) 522-89-97, e-mail: skarga_bandurova@ukr.net
ORCID ID: 0000-0003-3458-8730

УДК 004.832.34+303.442.3

Комплексний підхід до обробки різнорідних медичних даних з відсутніми значеннями / Білобородова Т.О., Скарга-Бандурова І.С. // Вісник НТУ "ХПІ". Серія: Інформатика та моделювання. – Харків: НТУ "ХПІ". – 2018. – № 42 (1318). – С. 180 – 187.

Показана необхідність врахування змішаних наборів даних з відсутніми значеннями. Запропоновано узагальнений підхід до обробки різнорідних медичних даних з урахуванням типу даних, механізму їх відсутності. Проведене опрацювання трьох різних наборів даних з використанням запропонованого підходу та контрольної моделі. Надано порівняння ефективності опрацювання даних. Іл.: 2. Табл.: 2. Бібліогр.: 12 назв.

Ключові слова: різнорідні медичні дані; відсутні значення; обробка; контрольна модель.

УДК 004.832.34+ 303.442.3

Комплексный подход к обработке разнородных медицинских данных с отсутствующими значениями / Белобородова Т.А., Скарга-Бандурова И.С. // Вестник НТУ "ХПИ". Серия: Информатика и моделирование. – Харьков: НТУ "ХПИ". – 2018. – № 42 (1318). – С. 180 – 187.

Показана необходимость учета смешанных наборов данных с отсутствующими значениями. Предложен обобщенный подход к обработке разнородных медицинских данных с учетом типа данных, механизма их отсутствия. Проведена обработка трех различных наборов данных с использованием предложенного подхода и контрольной модели. Предоставлено сравнения эффективности обработки данных. Ил.: 2. Табл.: 2. Библиогр.: 12 назв.

Ключевые слова: разнородные медицинские данные; отсутсвующие значения; обработка; контрольная модель.

UDC 004.67:618.3

A Comprehensive Approach for Processing Heterogeneous Medical Data with Missing Values / Biloborodova T.O., Skarga-Bandurova I.S. // Herald of the National Technical University "KhPI". Series of "Informatics and Modeling". – Kharkov: NTU "KhPI". – 2018. – №.42 (1318). – P. 180 – 187.

The need to take into account mixed data sets with missing values is shown. A generalized approach to the processing of heterogeneous medical data is proposed, taking into account the type of data, the mechanism of their absence. Three different sets of data have been processed using the proposed approach and control model. The comparison of the efficiency of data processing is given. Figs.: 2. Tabl.: 2, Refs.: 12 titles.

Keywords: heterogeneous medical data; missing values; data processing; control model.