УДК 004:616-094:519.233.3
DOI: https://doi.org/10.11603/mie.1996-1960.2022.3.13359

ДИСКРИМІНАЦІЯ ДАНИХ У ПАТОМОРФОЛОГІЇ. ШЛЯХИ ПОДОЛАННЯ

О. П. Мінцер, Н. О. Сінєнко[1]

Національний університет охорони здоров'я України імені П. Л. Шупика
[1]Міністерство охорони здоров'я України

Розглянуто питання сучасної аналітики, серед яких найгострішим є ризик дискримінації даних. Дослідження спрямовано на виявлення причин і наслідків дискримінації в інтелектуальному аналізі даних, а також на вивчення потенційних рішень цієї проблеми в патоморфології. Для встановлення складних взаємозв'язків між якістю оригінального зразка, якістю сканованого зображення, кількісними характеристиками застосованого стиснення, впливом на діагностичну інтерпретацію та інших пов'язаних діагностичних процедур (структурна класифікація, прогноз тощо) потрібно виконати значний обсяг робіт. Запропоновано алгоритм 4S (systematization, structuring stability of states), пов'язаний із використанням технології створення стійких морфологічних, гістологічних чи інших структур. Досвід його використання дає основу для обережного оптимізму.

Ключові слова: дискримінація даних, великі дані, стиснення даних, віртуальні слайди патології, модель візуальної дискримінації, нейронні згорткові мережі, алгоритм 4S.

DATA DISCRIMINATION IN PATHOMORPHOLOGY. WAYS OF COPING

O. P. Mintser, N. O. Sinyenko[1]

Shupyk National Healthcare University of Ukraine
[1]Ministry of Healthcare of Ukraine

Background. The issues of modern data analytics are considered, among which the risk of discrimination is the most acute. There are still no major studies on this topic in pathomorphology. Our work aims to identify the causes and consequences of discrimination in data mining and explore potential solutions to this problem. The purpose of the work was to identify the main reasons for the appearance of data discrimination in pathomorphology and conceptual substantiation of ways to overcome incomplete or biased data.

Materials and methods. Results. There are a large number of available compression schemes that help to solve various artifact problems (for example, blocking and color distortion, which is especially important for pathology, hot spot selection), but do not solve them all. Does not have a definitive solution to the effect of compressing the initial scan. To date, there is too little research on this topic.

Conclusions. A significant amount of work needs to be done before the complex relationships between the quality of the original sample, the quality of the scanned image, the quantitative characteristics of the applied compression and the impact on diagnostic interpretation and other related diagnostic procedures (structural classification, prognosis etc.). The proposed 4S algorithm (systematization, structuring stability of states), associated with the use of technology for creating stable morphological, histological, or other structures. The experience of its use provides a basis for cautious optimism.

Keywords: data discrimination, big data, data compression, virtual pathology slides, visual discrimination model, neural convolutional networks, 4S algorithm.

Вступ. Дискримінація даних, що також називається «дискримінацією за алгоритмом», представляє собою упередженість, що виникає, коли попередньо визначені типи даних або джерела даних заздалегідь або непередбачено обробляються інакше, ніж інші.

Дискримінація даних може призвести, перш за все, до суттєвого зниження якості надання медичної допомоги, недостатньої діагностики захворювань та прогнозування результатів їх лікування. Якісний догляд вимагає спільної комунікації, обміну інформацією та прийняття координованих рішень між лікарями та пацієнтами. Повні та точні дані про пацієнтів особливо важливі при обробленні великих даних, що використовуються для створення інструментів підтримки прийняття клінічних рішень та інформування ініціатив у галузі точної медицини. Однак систематично відсутні дані можуть спотворювати ці інструменти та загрожувати їх ефективності [1, 6, 9]. Дискримінація обумовлює зниження довіри пацієнта та його готовність ділитися інформацією зі своїм лікарем. Це, в свою чергу, має важливі наслідки для якості даних, доступних для прийняття медичних рішень та надання медичної допомоги. Пацієнти, які стикаються з дискримінацією даних, можуть

© О. П. Мінцер, Н. О. Сінєнко

опинитися в більш несприятливому положенні через систематичну відсутність даних у їх медичних картах. Велике значення має дискримінація даних у патоморфологічних дослідженнях [4].

Мета дослідження: виявлення основних причин появи дискримінації даних у патоморфології та концептуальне обґрунтування шляхів подолання неповних або упереджених даних.

Результати та їх обговорення. Найбільш часто дискримінація даних зустрічається при візуалізації зображень, особливо в телепатології. Надзвичайно великий розмір оцифрованих «віртуальних» слайдів, для зберігання яких можуть знадобитися гігабайти пам'яті, викликають значні затримки в передаванні даних для віддаленої інтерпретації зображень та інтерактивної візуалізації патологоанатомами [4]. Типовий слайд містить 1-4 см тканини, й одне сканування об'єктива зі збільшенням 40X при системному скануванні зі швидкістю 0,47 мкм /піксель дає 1-5 ГБ стиснутих даних зображення [8]. Ці вимоги можуть збільшуватися залежно від клінічного завдання. У деяких ситуаціях потрібне сканування лише з низькою роздільною здатністю (об'єктив 20X), але в інших потрібна значно більша роздільна здатність (об'єктив 80X або 100X). Наприклад, DMetrix Ultrarapid Virtual Slide Processor має масив із 80 об'єктів із високою роздільною здатністю, що дає близько 12 ГБ відсканованих даних [8].

Використання «середніх значень» має свої ризики в процесі дискримінації даних. Робота із середніми значеннями віддає пріоритет узагальненому «ідеальному» представнику, що в найкращому випадку може відображати лише деякі з пріоритетів користувача. Зменшення рівня використання середніх значень усуває ризик того, що алгоритми розпізнавання патологоанатомічної картини навчаються на одній групі та неналежним чином використовуються для вимірювання іншої.

Алгоритми для прийняття рішень про дані мають постійно самонавчатися: завдяки використанню штучного інтелекту (ШІ) можливо реагувати на особливості кожного окремого випадку навіть у режимі реального часу. Здатність масштабувати індивідуальні особистості кожного випадку в режимі реального часу також можливе лише за допомогою ШІ. Ручне втручання просто не може забезпечити обслуговування тисячі користувачів одночасно, кожен у свій спосіб.

Шляхи подолання. Одним із способів роботи з величезними обсягами даних віртуальних слайдів є стиснення. Воно може зменшити величезний обсяг даних віртуального слайду, але оборотні (без втрат) методи обмежують скорочення даних менш ніж на 50 %, у той час як стиснення із втратами може погіршити якість зображення та точність діагностики. Проблема в тому, що важко (якщо взагалі можливо) визначити єдиний «мінімальний» рівень стиснення (а отже, й якість зображення) для використання в усіх клінічних питаннях [10]. На сьогоднішній день більшість досліджень, що вивчають стиснення даних у телепатології, проводилися з використанням JPEG, а метод JPEG-2000 до останніх років визнається кращим.

Важливим удосконаленням в останні роки стала поява методів стиснення, адаптованих до людського сприйняття. З клінічної точки зору розроблення методів стиснення, адаптованих до людського сприйняття, може мати дві істотні переваги. По-перше, це може призвести до кращого зберігання найважливішої інформації про зображення зі значним зниженням вимог до зберігання. По-друге, інтелектуальне стиснення та швидке передавання даних зображень можуть забезпечити перегляд патоморфологічного матеріалу зі значним підвищенням продуктивності патологоанатомів, які використовують цифрові зображення замість предметного скла. Запропонований метод може допомогти усунути основні перешкоди для більш широкого впровадження технології цифрового оброблення зображень у патології, а саме труднощі до довгострокового зберігання зображень та поточний стан технології, що знижує продуктивність.

Власне кажучи, це й було показано в дослідженні [3]. Показана корисність моделі візуальної дискримінації (VDM) та інших показників спотворення для прогнозування швидкості передавання даних JPEG 2000, що відповідає стиску без візуальних втрат віртуальних слайдів для зразків біопсії молочної залози.

Інший підхід був розроблений для автоматизованого аналізу зображень під час кількісно вимірюваного стеатозу на гістологічних зображеннях [2]. Підкреслено, що просторова неоднорідність стеатозу може зробити його кількісне оцінювання ненадійним. Аналіз гарячих точок є найпоширенішим підходом до кількісного оцінювання неоднорідно розподілених параметрів тканини. Ідея полягає в тому, щоб розглядати лише області з особливо високими чи аномальними значеннями, які вважаються характерними для розподілу параметрів [5]. Зрозуміло, що його

результати критично залежать як від розташування, так і від розмірів областей, які розглядаються. При виконанні вручну вибір як розташування, так і розміру областей гарячих точок має тенденцію бути дуже суб'єктивним [7]. Сфокусовані оцінки області стеатозної тканини використовують концепцію аналізу «гарячих» точок на основі плиток для об'єктивного обчислення статистики. Оцінювалися сфокусовані бали за трьома наборами даних зображень зрізів печінки гризунів, що демонструють різний рівень стеатозу, спричиненого харчуванням. Було проведено стандартне оцінювання стеатозу, розраховане за допомогою більшості методів аналізу зображень.

Останнім часом у комп'ютерній графіці розповсюдження отримали тайли – зображення, що використовується для створення текстур шляхом укладання копій цього зображення на зразок «черепиці», так, щоб місця стику були непомітні. В тайловій графіці застосовується набір тайлів (тайлсет) і матриця клітинок, що визначає який тайл яку клітинку займатиме. Підходи з урахуванням тайлів роблять аналіз гарячих точок об'єктивнішим. Вони розбивають зображення тканин на регулярну сітку тайлів і визначають параметри тканини для кожного тайлу індивідуально. Обидві операції виконуються автоматичним аналізом зображення. Це дозволяє об'єктивно обирати області гарячих точок на основі значень параметрів тайлів. Крім того, тайл дозволяє охарактеризувати просторовий розподіл параметра тканини за допомогою статистики про його значення за осередками.

Добре відомий досвід використання обчислювальних методів, заснованих на нейронних згорткових мережах (Convolutional Neural Networks - CNN). Створений автономний конвеєр для ефективної класифікації різних гістопатологічних зображень для різних типів раку та продемонстрована корисність конвеєра для диференціальної діагностики двох підтипів раку легень, чотирьох біомаркерів раку сечового міхура та п'яти біомаркерів раку молочної залози. Крім того, ефективно був застосований конвеєр для розрізнення чотирьох показників імуногістохімічного фарбування раку сечового міхура та молочної залози.

Отже, існує велика кількість доступних схем стиснення, що вирішують, у тій або іншій мірі сприяють вирішенню різних проблем артефактів (наприклад, блокування та спотворення кольору, що особливо важливо для патології, вибір «гарячих» точок), але не вирішує їх усіх. Не має кінцевого рішення про вплив стиснення початкового сканування. На сьогоднішній день досліджень на цю тему надто мало.

В наших дослідження застосовано оригінальній алгоритм 4S (systematization, structuring stability of states), пов'язаний із використанням технології створення стійких станів морфологічних, гістологічних чи інших структур. Останні мають логічне поєднувати первинні елементи та полегшувати ефективне використання, збереження та спільне використання даних. Іншими словами створюються формальні моделі даних, що описують спосіб їх організації. Структури даних є «будівельними блоками» для складніших додатків. Вони розробляються шляхом об'єднання елементів даних у вторинну структуру метаданих, що є, взагалі, абстрактним типом даних, який має відношення лише до алгоритмів оброблення чи додатків. Поки що маємо відносно невеликий обсяг спостережень, що дає основу лише для обережного оптимізму.

Висновки. 1. Потрібно виконати значний обсяг роботи, перш, ніж будуть встановлені складні взаємозв'язки між якістю оригінального зразка, якістю сканованого зображення, кількісними характеристиками застосованого стиснення, впливом на діагностичну інтерпретацію та інші пов'язані діагностичні процедури (структурна класифікація, прогноз тощо).

2. Запропоновано алгоритм 4S, пов'язаний із використанням технології створення стійких морфологічних, гістологічних чи інших структур. Досвід його використання дає основу для обережного оптимізму.

Література.

1. Big Data and discrimination: perils, promises and solutions. A systematic review / Favaretto M., De Clercq E., Elger B. S. // J Big Data. – 2019. – № 6. – С. 12.

2. Focused scores enable reliable discrimination of small differences in steatosis / Homeyer A., Hammad S., Schwen L.O. et al. // Diagn Pathol. – 2018. – № 3. – С. 76.

3. Using a visual discrimination model for the detection of compression artifacts in virtual pathology images / Johnson J. P., Krupinski E. A., Yan M., Roehrig H. et al. // IEEE Trans Med Imaging. – 2011. – Vol. 30, № 2. – P. 306-314.

4. Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images / Khosravi P., Kazemi E., Imielinski M., Elemento O., Hajirasouliha I. // EBioMedicine. – 2018. – № 27. – P. 317-328.

5. Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome / O'Connor J. P. B., Rose C. J., Waterton J. C., Carano R. A. D. et al. // Clin Cancer Res. – 2015. – № 21. – P. 249-257.

6. Discrimination, trust, and withholding information from providers: Implications for missing data and inequity / Nong P., Williamson A., Anthony D., Platt J., Kardia Sh. // SSM – Population Health. – 2022. – Vol. 18, № 101092. – P. 1-7.

7. A methodology for comprehensive breast cancer Ki67 labeling index with intra-tumor heterogeneity appraisal based on hexagonal tiling of digital image analysis data / Plancoulaine B., Laurinaviciene A., Herlin P. // Virchows Arch. – 2015. – № 467. – P. 711-722.

8. An array microscope for ultrarapid virtual slide processing and telepathology. Design, fabrication, and validation study / Weinstein R.S., Descour M.R., Liang C. et al. // Human Pathol. – 2004. – № 35. – P. 1303-1314.

9. Нові тенденції розвитку систем представлення та управління даними. аналітичний погляд / Мінцер О. П., Бабінцева Л. Ю. // Медична інформатика та інженерія. – 2022. – № 1-2 (57-58). – С. 5-13.

10. Digital imaging in pathology: The case for standardization / Yagi Y., Gilberson J. R. // J Telemed Telecare. – 2005. – № 11. – P. 109-116.

References.

1. Favaretto, M., De Clercq, E., Elger, B. S. (2019). Big Data and discrimination: perils, promises and solutions. A systematic review. J Big Data, 6(12). DOI:10.1186/s40537-019-0177-4.

2. Homeyer, A., Hammad, S., Schwen, L. O., Dahmen, U., Höfener, H. et al. (2018). Focused scores enable reliable discrimination of small differences in steatosis. Diagn Pathol, 3(76). DOI:10.1186/s13000-018-0753-5.

3. Johnson, J. P., Krupinski, E. A., Yan, M., Roehrig, H. et al. (2011). Using a visual discrimination model for the detection of compression artifacts in virtual pathology images. IEEE Trans Med Imaging, 30(2), 306-314. DOI:10.1109/TMI.2010.2077308-2.

4. Khosravi, P., Kazemi, E., Imielinski, M., Elemento, O., Hajirasouliha, I. (2018). Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images. EBioMedicine, 27, 317-328. DOI:10.1016/j.ebiom.2017.12.026.

5. O'Connor, J. P. B., Rose, C. J., Waterton, J. C. et al. (2015). Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. Clin Cancer Res, 21, 49-57.

6. Nong, P., Williamson, A., Anthony, D., Platt, J., Kardia Sh. (2022). Discrimination, trust, and withholding information from providers: Implications for missing data and inequity. SSM - Population Health. 18(101092), 1-7.

7. Plancoulaine, B., Laurinaviciene, A., Herlin, P. (2015). A methodology for comprehensive breast cancer Ki67 labeling index with intra-tumor heterogeneity appraisal based on hexagonal tiling of digital image analysis data. Virchows Arch, 467, 711-722.

8. Weinstein, R. S., Descour, M. R., Liang, C., Barker, G., Scott, K. M. et al. (2004). An array microscope for ultrarapid virtual slide processing and telepathology. Design, fabrication, and validation study. Human Pathol, 35, 1303-1314.

9. Mintser O. P., Babintseva L. Yu. (2022). Novi tendentsiyi rozvytku system predstavlennya ta upravlinnya danymy. analitychnyy pohlyad. [New trends in the development of data presentation and management systems. Analytical view]. Medychna informatyka ta inzheneriia. [Medical informatics and engineering], 1-2 (57-58), 5-13. DOI: 10.11603/mie.1996-1960.2022.1-2.13104. [In Ukrainian].

10. Yagi, Y., Gilberson, J. R. (2005). Digital imaging in pathology: The case for standardization. J Telemed Telecare, 11, 109-116.

ORCID:
Ozar P. Mintser: 0000-0002-7224-4886
Natalia O. Sinyenko: 0000-0002-8013-3449