nature

CAREER COLUMN | 01 February 2024

In the AI science boom, beware: your results are only as good as your data

Machine-learning systems are voracious data consumers – but trustworthy results require more vetting both before and after publication.

By Hunter Moseley

Hunter Moseley says that good reproducibility practices are essential to fully harness the potential of big data. Credit: Hunter N.B. Moseley

We are in the middle of a data-driven science boom. Huge, complex data sets, often with large numbers of individually measured and annotated 'features', are fodder for voracious artificial intelligence (AI) and machine-learning systems, with details of new applications being published almost daily.

But publication in itself is not synonymous with factuality. Just because a paper, method or data set is published does not mean that it is correct and free from mistakes. Without checking for accuracy and validity before using these resources, scientists will surely encounter errors. In fact, they already have.

In the past few months, members of our bioinformatics and systems-biology laboratory have reviewed state-of-the-art machine-learning methods for predicting the metabolic pathways that metabolites belong to, on the basis of the molecules' chemical structures[1]. We wanted to find, implement and potentially improve the best methods for identifying how metabolic pathways are perturbed under different conditions: for instance, in diseased versus normal tissues.

We found several papers, published between 2011 and 2022, that demonstrated the application of different machine-learning methods to a gold-standard metabolite data set derived from the Kyoto Encyclopedia of Genes and Genomes (KEGG), which is maintained at Kyoto University in Japan. We expected the algorithms to improve over time, and saw just that: newer methods performed better than older ones did. But were those improvements real?

Data leaks

Scientific reproducibility enables careful vetting of data and results by peer reviewers as well as by other research groups, especially when the data set is used in new applications. Fortunately, in keeping with best practices for computational reproducibility, two of the papers[2,3] in our analysis included everything that is needed to put their observations to the test: the data set they used, the computer code they wrote to implement their methods and the results generated from that code. Three of the papers[2–4] used the same data set, which allowed us to make direct comparisons. When we did so, we found something unexpected.

It is common practice in machine learning to split a data set in two and to use one subset to train a model and another to evaluate its performance. If there is no overlap between the training and testing subsets, performance in the testing phase will reflect how well the model learns and performs. But in the papers we analysed, we identified a catastrophic 'data leakage' problem: the two subsets were cross-contaminated, muddying the ideal separation. More than 1,700 of 6,648 entries from the KEGG COMPOUND database — about one-quarter of the total data set — were represented more than once, corrupting the cross-validation steps.

RELATED

NatureTech

When we removed the duplicates in the data set and applied the published methods again, the observed performance was less impressive than it had first seemed. There was a substantial drop in the F_1 score — a machine-learning evaluation metric that is similar to accuracy but is calculated in terms of precision and recall — from 0.94 to 0.82. A score of 0.94 is reasonably high and indicates that the algorithm is usable in many scientific applications. A score of 0.82, however, suggests that it can be useful, but only for certain applications — and only if handled appropriately.

It is, of course, unfortunate that these studies were published with flawed results stemming from the corrupted data set; our work calls their findings into question. But because the authors of two of the studies followed best practices in computational scientific reproducibility and made their data, code and results fully available, the scientific method worked as intended, and the flawed results were detected and (to the best of our knowledge) are being corrected.

The third team, as far as we can tell, included neither their data set nor their code, making it impossible for us to properly evaluate their results. If all of the groups had neglected to make their data and code available, this data-leakage problem would have been almost impossible to catch. That would be a problem not just for the studies that were already published, but also for every other scientist who might want to use that data set for their own work.

More insidiously, the erroneously high performance reported in these papers could dissuade others from attempting to improve on the published methods, because they would incorrectly find their own algorithms lacking by comparison. Equally troubling, it could also complicate journal publication, because demonstrating improvement is often a requirement for successful review — potentially holding back research for years.

Encouraging reproducibility

So, what should we do with these erroneous studies? Some would argue that they should be retracted. We would caution against such a knee-jerk reaction — at least as a blanket policy. Because two of the three papers in our analysis included the data, code and full results, we could evaluate their findings and flag the problematic data set. On one hand, that behaviour should be encouraged — for instance, by allowing the authors to publish corrections. On the other, retracting studies with both highly flawed results and little or no support for reproducible research would send the message that scientific reproducibility is not optional. Furthermore, demonstrating support for full scientific reproducibility provides a clear litmus test for journals to use when deciding between correction and retraction.

Now, scientific data are growing more complex every day. Data sets used in complex analyses, especially those involving AI, are part of the scientific record. They should be made available — along with the code with which to analyse them — either as supplemental material or through open data repositories, such as Figshare (Figshare has partnered with Springer Nature, which publishes *Nature*, to facilitate data sharing in published manuscripts) and Zenodo, that can ensure data persistence and provenance. But those steps will help only if researchers also learn to treat published data with some scepticism, if only to avoid repeating others' mistakes.

doi: https://doi.org/10.1038/d41586-024-00306-2

This is an article from the Nature Careers Community, a place for Nature readers to share their professional experiences and advice. Guest posts are encouraged.

References

1. Huckvale, E. D. & Moseley, H. Preprint at biorXiv https://doi.org/10.1101/2023.10.03.560711 (2023).

2. Baranwal, M. *et al. Bioinformatics* **36**, 2547–2553 (2020).

3. Yang, Z., Liu, J., Wang, Z., Wang, Y. & Feng, J. In *2020 IEEE International Conference on Bioinformatics and Biomedicine* 126–131 (Institute of Electrical and Electronics Engineers, 2020).

4. Du, B.-X. *et al. Bioinformatics* **38**, i325–i332 (2022).

COMPETING INTERESTS

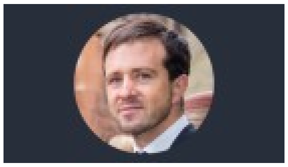
The author declares no competing interests.

Latest on:

Machine learning Databases Research data

DeepLabCut: the motion-tracking tool that went viral

TECHNOLOGY FEATURE | 20 MAY 24



Why mathematics is set to be revolutionized by AI

WORLD VIEW | 14 MAY 24

How does ChatGPT 'think'? Psychology and neuroscience crack open AI large language models

NEWS FEATURE | 14 MAY 24

Nature (*Nature*) | ISSN 1476-4687 (online) | ISSN 0028-0836 (print)