Machine Learning: Addressing the Limitations of Real-World Data

Andrew Cox, PhD Research Scientist, Retrospective Observational Studies
Joseph Lee, PhD Senior Research Associate, Modeling & Simulation

The field of retrospective observational studies (ROS) and real-world data (RWD) is undergoing fast-paced development. The increase in available data sources has been accompanied by a rapid methodological evolution to address problems commonly encountered in RWD studies. For example, problems may include the evidence being fragmented across multiple datasets; the target population being inconsistently identified or present in the data but undiagnosed; or the information only being available in free-text elements of the data source. Here, we describe how an approach called machine learning can help you solve these types of problems and get the most out of your data.

What is machine learning?

At its heart, machine learning is an algorithmic approach to extract meaning from data. Although you may not have heard of it, you encounter machine learning every day without realizing it. It is used by email providers to filter spam from your email, by banks to prevent credit card fraud, and by companies like Amazon, Google, and Facebook to present content personalized to your interests. In short, machine learning permeates everyday life.

Machine learning originated from multiple fields, including classical statistics, computer science, artificial intelligence, and data mining. It is most often compared with classical statistics, but there are a few key differences worth noting. Classical statistics generally assumes the data is generated by an underlying probability model. It is largely concerned with hypothesis testing, goodness of fit testing, and inference from historical data. In contrast, machine learning assumes the data is generated by an unknown mechanism and is largely concerned with learning the patterns within data to make accurate predictions.

Machine learning methods

Although machine learning is ubiquitous outside of biomedical fields, it has not yet seen the same level of adoption in the healthcare industry. However, an increase in the use of machine learning is being seen as the volume and variety of data grows.

Machine learning methods are well-suited to large datasets that incorporate a wide variety of data types, including unstructured data like text, CT scans, or genomic data. There are many different machine learning techniques with catchy and enigmatic names: classification and regression tree (CART), random forest, AdaBoost, support vector machines (SVM), neural nets, Bayesian networks, and C4.5 are frequently used. Some techniques excel in specialized applications; other times choosing the right technique is more a matter of art than science.

Within the biomedical field, there is a great need for transparent and human interpretable output. Decision makers must trust the model to act on its results. In this scenario, an easy-to-explain method such as a decision tree may be preferable to an opaque method such as a neural network, even if the decision tree is less powerful. Decision trees produce visible rules that are easy to follow and understand, meaning people like front-line clinicians can quickly apply it and communicate its results.

Machine learning methods are well-suited to large datasets that incorporate a wide variety of data types...

Use cases

Machine learning has many applications in the healthcare industry. In a straightforward translation of traditional business analytics, machine learning can be used to predict which patients will discontinue a drug for a chronic disease. A company can then take action to reduce patient "churn" and increase revenue.

Another important area for machine learning is improving diagnoses and predicting disease outcomes. It has been used for the early detection of Alzheimer's disease[1] and can improve the accuracy of cancer outcome predictions by 15–20%.[2]

Machine learning is an ideal approach to solve problems presented by retrospective and real-world data. For instance, in a recent study done at Evidera, the prevalence of post-stroke spasticity (PSS) was observed to be as low as 1% in clinical practice research datalink (CPRD) data. This observation was well below the 20–30% prevalence of PSS in the published literature. The speculation was that not all stroke patients that developed spasticity were given a diagnostic code for spasticity by their primary care physician. Such an underreporting of PSS in CPRD data would make any future studies of costs of care subject to bias.

If you proceeded to conduct this study using the limited number of PSS patient records, your results would be biased because the few cases identified in the data were likely to be a subpopulation of the most severe cases. To overcome this limitation, we used our expertise in machine learning to identify a previously undiagnosed population of PSS patients in the CPRD dataset. With the help of key opinion leaders who helped create a list of treatments frequently used for PSS, we boosted the sample size from 665 to nearly 4,000 PSS patients and reduced the bias in results when compared to conducting analyses only on the 665 patients who received a diagnostic code. This study is a perfect example of how machine learning can overcome the perceived limitations of RWD and yield considerable benefits.

Limitations

Machine learning techniques can be immensely powerful, but they require careful and expert application. Special care must be taken to avoid "overfitting," in which the model produces highly accurate predictions for the data it was trained on but is completely ineffective and inaccurate when used to make predictions on new data. Overfitting is one of the most frequent mistakes seen in studies using these techniques.

Machine learning is an ideal approach to solve problems presented by retrospective and real-world data.

Conclusion

The success of machine learning techniques led to its rapid and widespread adoption across a diverse range of fields and disciplines. Since skills and expertise in machine learning are still rare in the healthcare industry, few realize that it can be used to solve many of the problems frequently presented by RWD studies. As the benefits of machine learning are better understood, we are likely to see a large increase in its usage over the coming years.

For more information, please contact Andrew.Cox@evidera.com or Joseph.Lee@evidera.com.

REFERENCES

[1] Lebedev AV, Westman E, Van Westen GJ, et al. Random Forest Ensembles for Detection and Prediction of Alzheimer's Disease with a Good Between-Cohort Robustness. *NeuroImage Clin*. 2014 Aug 28; 6:115-125.

[2] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine Learning Applications in Cancer Prognosis and Prediction. *Comput Struct Biotechnol J*. 2014 Nov 15; 13:8-17.