Machine-Learning Algorithms Can Help Health Care Litigation

By Lisa B. Pinheiro, Jimmy Royer, Nick Dadson and Paul E. Greenberg; Analysis Group, Inc.

Law360, New York (June 8, 2016, 1:10 PM ET)

Lisa B. Pinheiro

Jimmy Royer

Nick Dadson

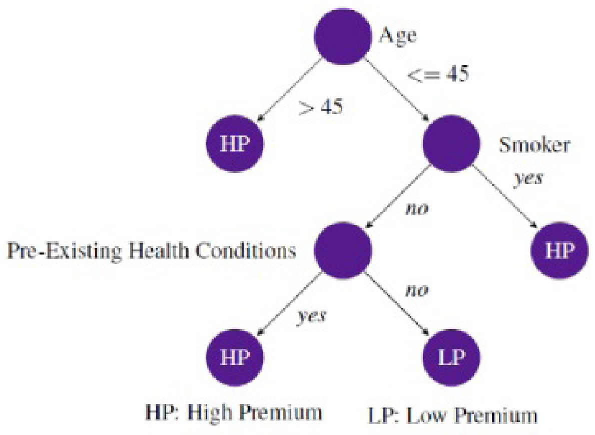
Paul E. Greenberg

Machine-learning algorithms are ubiquitous these days. Technology giants like Netflix Inc., Amazon.com Inc. and Google Inc. use them to suggest items customers might like based on their past browsing. Scientists use them to identify gene mutations associated with treatment resistance or amenable to targeted drug therapy. And doctors use them for image classification, early disease detection and better treatment outcomes. These algorithms can improve quality of life and can even help save lives. But what are machine-learning algorithms and what is their potential role in health care litigation?

To answer this question, let's start from the beginning: the algorithm. An algorithm is a sequence of step-by-step instructions for solving a particular problem. For example, a health insurance provider can, in principle, ask applicants a series of questions to determine whether a high- or low-premium policy would be more appropriate. This process can be represented by a decision tree, as shown in the figure. In this simplified algorithm, the applicant is classified as either a low-premium or high-premium customer depending on his answers to a few questions.

Algorithms are used in our everyday lives to replace costly or time-consuming repetitive tasks, such as organizing data in alphabetical order or using a search and replace function in a document. The health insurance premium example of course vastly oversimplifies the insurer's actual problem as it likely has access to a lot more information about

Simple Decision Tree Algorithm for Health Insurance Premiums



its clients and may be constrained by law in how it goes about distinguishing customer types in this context. Furthermore, determining the most important predictors of future claims and how to gather the most relevant information about potential claimants is not a simple problem. This is where machine learning can come into play.

Machine-learning algorithms are used to detect complex, often unforeseen patterns within rich datasets. There are two general categories of algorithms: unsupervised and supervised. Unsupervised machine-learning algorithms are typically used to group large amounts of data into categories, where the categories are not specified in advance. In such cases, the researcher relies on the algorithm to identify patterns in the data beyond what otherwise may look like unstructured noise. This type of algorithm can be used for clustering populations into different subgroups for further analysis (e.g., segmenting patients based on patterns of biomarkers identified by the algorithm). Unsupervised algorithms can be used to generate hypotheses, and thus, often precede use of a supervised algorithm. Supervised machine-learning algorithms start out with a hypothesis and categories that are set out in advance. These algorithms are then "trained" on data for which the outcomes of interest are known, with the training process continuing until a desired level of accuracy is achieved. These results are then used to make predictions based on out-of-sample data for which the outcome of interest is not known.

Machine-learning algorithms may sound very similar to more familiar statistical models; indeed most statistical models can be viewed as a form of machine-learning algorithm. However, widely-used statistical models such as regressions often impose a linear relationship between the variables on the one hand and the outcome of interest on the other. This simplification allows for easy interpretation of the relative impact of each variable. More advanced machine-learning algorithms impose much less structure and can therefore detect very complex and intricate relationships in high-dimensional data (i.e., data with several different types of variables, possibly including quantitative, text and image information). Advances are now being made in analyzing the output of these algorithms to permit assessment of the relative importance of each variable, as can already be done with linear regressions[1].

In the litigation context, machine learning generally refers to the supervised version and there is a wide array of potential uses. E-discovery is a classic example. As technological innovation has made the storage of information extremely simple and inexpensive, the amount of information that litigants need to sift through continues to increase significantly. But determining which documents or emails are "relevant" can quickly become a formidable task. Automating this process using keyword searches can save time and money but often produces poor quality results. Machine-learning algorithms can be trained on a sample of documents where the outcome of interest is known (i.e., the relevance of the document to the litigation) to substantially reduce the efforts of human reviewers on all the remaining documents where that outcome is not known[2].

Fraud detection is another application of machine-learning algorithms that can be easily extended to a health care litigation setting. Examples of familiar fraud allegations include filing false medical claims for unused services, faking eligibility documentation to obtain lower premiums and claiming benefits illegally using another person's coverage[3]. The quantity and complexity of health insurance data make it extremely cumbersome and expensive to detect these behaviors with other methods[4]. In addition to reducing the time spent on an otherwise tedious process, machine-learning algorithms can output a "fraud likelihood" for each case, which means that resources can be allocated to suspicious cases based on preset likelihood thresholds[5, 6]. These same algorithms can also be used to help measure potential exposure in a litigation context.

Another potential application of machine learning in the health care litigation setting involves prediction of counterfactual scenarios. For example, litigation cases involving transfer pricing disputes, off-label promotion or kickback allegations, or disputes related to joint-marketing agreements often focus on the sales impact of marketing. One question that arises in such cases is how much of the drug in question would have been prescribed even in the absence of certain types of marketing efforts by a pharmaceutical manufacturer. As electronic medical records and other electronic de-identified patient data become more widely available and able to be combined with other data concerning physician characteristics, drug features, scientific findings and reimbursement policies, machine learning algorithms will offer a robust tool for predicting physician prescribing under alternative assumptions concerning the nature and extent of marketing.

Textual analysis of patent claims represents another area where conventional statistical methods are limited, but machine-learning algorithms (such as natural language processing) can be beneficial. Challenges to patents on branded drugs continue to rise, and given the high stakes and corresponding uncertainty associated with such challenges, machine learning could play a key role in how related litigation unfolds[7]. That is, machine-learning algorithms could be used to predict the likelihood of a successful patent challenge (e.g., whether an obviousness claim will be sustained) and thus could help the parties determine whether to engage in litigation, and if so, at what point to negotiate a settlement versus continue to litigate. Whereas standard statistical analyses would be of limited value as they can incorporate attention only to relatively few variables in this context, the predictive power of machine-learning algorithms is greatly increased with benefit of much more information. Such information includes, for instance, the text of the patent claims as well as that of all other patents in its class, together with numerous other variables, such as patent reviewer characteristics, forward citations, parties involved and the particular court. In this sense, the algorithm is a learning process that accounts for all the available information in increasingly accurate ways. In so doing, it approximates the actual process that leads to a final judgment at the patent office.

These are only a few examples among many. Machine-learning algorithms have a widearray of potential uses in the health care litigation context. Their appeal has largely been due to their ability to process big data — datasets that are too large and too complex for traditional statistical approaches. Machine-learning algorithms made a name for themselves initially in web-based industries, featured in services we now view as commonplace. Other industries have taken note and we now see machine-learning algorithms used across a variety of industries including health care.

In summary, supervised machine-learning algorithms are most likely to add value when three conditions are met:

1) The goal of the analysis is to *predict* an outcome;
2) *Out-of-sample* performance is the desired measure of success; and
3) A rich dataset is available to take advantage of interactions among many poten tial predictors with a complex inter-relationship (e.g., likely a nonlinear function of many factors but difficult to specify its form in advance).

The examples above meet these conditions. They aim to predict whether emails are relevant to a particular litigation, whether claims are fraudulent, what drug a physician will prescribe or whether patents will be successfully challenged in court. It is crucial to these examples that the patterns isolated in the known sample of data be generalizable to other samples (e.g., that the fraudulent claim detection algorithm perform well on rest of the sample). And, in all cases, there is a large and varied set of variables from which to draw.

If the litigation-related problem you face fits these characteristics, it is likely worthwhile to consider using machine-learning models. You may find insights to problems you thought were too complex or too time consuming to solve with traditional tools and approaches.

The opinions expressed are those of the author(s) and do not necessarily reflect the views of the firm, its clients, or Portfolio Media Inc., or any of its or their respective affiliates. This article is for general information purposes and is not intended to be and should not be taken as legal advice.

Endnotes

[1] Rose, Sherri. “Targeted Learning for Variable Importance.” Handbook of Big Data (2016): 411.

[2] Electronic Discovery (E-Discovery) has become a classic litigation application of text mining. See Gordon V. Cormack and Maura R. Grossman, “Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery,” in Proceedings of SIGIR 2014: The 37th Annual ACM SIGIR Conference on Research and Development in Information Retrieval (2014): 153-62. See also Ghayad et al., “Making the Most of Document Analytics,” Law360, Dec. 1, 2015.

[3] Fraud detection is important in other industries as well (e.g., credit cards, telecommunications, auto insurance, and online auctions). See Aisha Abdallah, Mohd Aizaini, and Anazida Zainal, “Fraud detection system: A survey,” Journal of Network and Computer Applications vol. 68 (2016): 90-113.

[4] The data used to detect health insurance fraud can come from a variety of sources including: insurance claims, subscriber characteristics (e.g., age, geographic location), and clinical information (e.g., medical treatment and prescription drug history). See Prern Dua and Sonali Bais, “Supervised Learning Methods for Fraud Detection in Healthcare Insurance,” in Machine Learning in Healthcare Informatics, ed. Sumeet Dua et al. (Berlin: Springer-Verlag, 2014), 265; Ilker Kose, Mehmet Gokturk, and Kemal Kilic, “An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance,” Applied Soft Computing vol. 36 (2015): 283-99.

[5] Aisha Abdallah, Mohd Aizaini, and Anazida Zainal, “Fraud detection system: A survey,” Journal of Network and Computer Applications 68 (2016): 90-113; Hussein Almuallim, Shigeo Kaneda, and Yasuhiro Akiba, “Development and Applications of Decision Trees,” in Expert Systems: The Technology of Knowledge Management and Decision Making for the 21st Century, ed. Cornelius T. Leondes (San Diego: Academic Press, 2002), 54.

[6] Shi et al., for example, use over 18,000 records from a U.S. insurance company with detailed information on policy types, claim/audit, policy producer, and client which include a label for potential fraudulent claims. They use these data to compare multiple different classification methods, such as the decision-tree algorithm described above. See Yong Shi et al. “Health Insurance Fraud Detection,” in Optimization Based Data Mining: Theory and Applications, ed. Yong Shi et al. (London: Springer-Verlag, 2011), 233-35.

[7] In 2011-2012, 80% of the NMEs experiencing first generic entry experienced a patent challenge; prior to 1998, this was less than 20%. Henry G. Grabowski, Genia Long, and Richard Mortimer, “Recent Trends in Brand Name and Generic Drug Competition,” Journal of Medical Economics vol. 7, no. 3 (2014): 207-14. For the change in rules, see Food and Drug Administration, “FDA Guidance for Industry When Multiple ANDAs are Submitted on the Same Day,” accessed June 2, 2016. Laura E. Panettoni, “The Effect of Paragraph IV Decisions and Generic Entry Before Patent Expiration on Brand Pharmaceutical Firms,” Journal of Health Economics vol. 30, 126-45.