Top 10 data mining mistakes

Avoid common pitfalls on the path to data mining success

JOHN ELDER,
ELDER RESEARCH INC.

Mining data to extract useful and enduring patterns is a skill arguably more art than science. Pressure enhances the appeal of early apparent results, but it's too easy to fool yourself. How can you resist the siren songs of the data and maintain an analysis discipline that will lead to robust results? What follows are the most common mistakes made in data mining. Note: The list was originally a Top 10, but after compiling the list, one basic problem remained – mining without proper data. So, numbering like a computer scientist (with an overflow problem), here are mistakes Zero to 10.

ZERO Lack proper data. To really make advances with an analysis, one must have labeled cases, such as an output variable, not just input variables. Even with an output variable, the most interesting type of observation is usually the most rare by orders of magnitude. The less probable the interesting events, the more data it takes to obtain enough to generalize a model to unseen cases. Some projects shouldn't proceed until enough critical data is gathered to make them worthwhile.

> The most exciting phrase in research is not the triumphal "Aha!" of discovery, but the puzzled uttering of "That's odd."

ONE Focus on training. Early machine learning work often sought to continue learning (refining and adding to the model) until achieving exact results on known data – which, at the least, insufficiently respects the incompleteness of our knowledge of a situation. Obsession with getting the most out of training cases focuses the model too much on the peculiarities of that data to the detriment of inducing general lessons that will apply to similar, but unseen, data. Try resampling, with multiple modeling experiments and different samples of the data, to illuminate the distribution of results. The mean of this distribution of evaluation results tends to be more accurate than a single experiment, and it also provides, in its standard deviation, a confidence measure.

TWO Rely on one technique. For many reasons, most researchers and practitioners focus too narrowly on one type of modeling technique. At the very least, be sure to compare any new and promising method against a stodgy conventional one. Using only one modeling method forces you to credit or blame it for the results, when most often the data is to blame. It's unusual for the particular modeling technique to

0 Lack proper data

1 Focus on training

2 Rely on one technique

3 Ask the wrong question

4 Listen (only) to the data

5 Accept leaks from the future

6 Discount pesky cases

7 Extrapolate

8 Answer every inquiry

9 Sample casually

10 Believe the best model

make more difference than the expertise of the practitioner or the inherent difficulty of the data. It's best to employ a handful of good tools. Once the data becomes useful, running another familiar algorithm, and analyzing its results, adds only 5-10 percent more effort.

THREE Ask the wrong question. It's important first to have the right project goal or ask the right question of the data. It's also essential to have an appropriate model goal. You want the computer to feel about the problem like you do – to share your multi-factor score function, just as stock grants give key employees a similar stake as owners in the fortunes of a company. Analysts and tool vendors, however, often use squared error as the criterion, rather than one tailored to the problem.

FOUR Listen (only) to the data. Inducing models from data has the virtue of looking at the data afresh, not constrained by old hypotheses. However, don't tune out received wisdom while letting the data speak. No modeling technology alone can correct for flaws in the data. It takes careful study of how the model works to understand its weakness. Experience has taught once brash analysts that those familiar with the domain are usually as vital to the solution as the technology brought to bear.

FIVE Accept leaks from the future. Take this example of a bank's neural network model developed to forecast interest rate changes. The model was 95 percent accurate – astonishing given the importance of such rates for much of the economy. Cautiously ecstatic, the bank sought a second opinion. It was found that a version of the output variable had accidentally been made a candidate input. Thus, the output could be thought of as only losing 5 percent of its information as it traversed the network. Data warehouses are built to hold the best information known to date; they are not naturally able to pull out what was known during the timeframe that you wish to study. So, when storing data for future mining, it's important to date-stamp records and to archive the full collection at regular intervals. Otherwise, it will be very difficult to recreate realistic information states, leading to wrong conclusions.

SIX Discount pesky cases. Outliers and leverage points can greatly affect summary results and cloud general trends. Don't dismiss them; they could be the result. When possible, visualize data to help decide whether outliers are mistakes or findings. The most exciting phrase in research is not the triumphal "Aha!" of discovery, but the puzzled uttering of "That's odd." To be surprised, one must have expectations. Make hypotheses of results before beginning experiments.

SEVEN Extrapolate. We tend to learn too much from our first few experiences with a technique or problem. Our brains are desperate to simplify things. Confronted with conflicting data, early hypotheses are hard to dethrone - we're naturally reluctant to unlearn things we've come to believe, even after an upstream error in our process is discovered. The antidote to retaining outdated stereotypes about our data is regular communication with colleagues

about the work, to uncover and organize the unconscious hypotheses guiding our explorations.

EIGHT Answer every inquiry. If only a model answered "Don't know!" for situations in which its training has no standing! Take the following example of a model that estimated rocket thrust using engine temperature, T, as an input. Responding to a query where T = 98.6 degrees provides ridiculous results, as the input, in this case, is far outside the model's training bounds. So, how do we know where the model is valid; that is, has enough data close to the query by which to make a useful decision? Start by noting whether the new point is outside the bounds, on any dimension, of the training data. But also pay attention to how far away the nearest known data points are.

NINE Sample casually. The interesting cases for many data mining problems are rare and the analytic challenge is akin to finding needles in a haystack. However, many algorithms don't perform well in practice, if the ratio of hay to needles is greater than about 10 to 1. To obtain a near-enough balance, one must either down-sample to remove most common cases or up-sample to duplicate rare cases. Yet it is a mistake to do either casually. A good strategy is to "shake before baking"; that is, to randomize the order of a file before sampling. Split data into sets first, then up-sample rare cases in training only. A stratified sample will often save you trouble. Always consider which variables need to be represented in each data subset and sample separately.

TEN Believe the best model. Don't read too much into models; it may do more harm than good. Too much attention can be paid to particular variables used by the best data mining model – which likely barely won out over hundreds of others of the millions (to billions) tried – using a score function only approximating your goals, and on finite data scarcely representing the underlying data-generating mechanism. Better to build several models and interpret the resulting distribution of variables, rather than the set chosen by the single best model.

How will we succeed?

Modern tools, and harder analytic challenges, mean we can now shoot ourselves in the foot with greater accuracy and more power than ever before. Success is improved by learning from experience; especially our mistakes. So go out and make mistakes early! Then do well, while doing good, with these powerful analytical tools. ■

[author bio]
John Elder, PhD, is the founder of Elder Research Inc., a leading data mining consulting firm (www.datamininglab.com). The material in this article is from Chapter 20 of the book, *Handbook of Statistical Analysis & Data Mining Applications*, with Bob Nisbet and Gary Miner.
Elder@datamininglab.com

online
Read full excerpt:
www.sas.com/sascom-excerpt

Buy the book:
www.sas.com/sascom-elderbook

THE PATH TO DATA MINING SUCCESS:

PERSISTENCE: Attack data mining problems from different angles; automate essential steps; perform resampling tests; and externally check your work.

ATTITUDE: An optimistic attitude can work wonders for results, especially in a team setting.

TEAMWORK: Business and statistical experts must cooperate closely and share the same goals to make the best progress and be successful.

HUMILITY: Ask the right questions, learn from others and step back to see the big picture.