Data Mining Misconceptions #1: The 50/50 Problem

By Tim Graettinger

This fall will mark my twentieth year as a data mining professional. Thank you. During that time, I worked at five different companies – mostly startups - and consulted for many, many clients. Changes to the data mining field during that period are startling, in terms of the computational horsepower available, the size of the databases being generated, and the software tools developed to model and analyze them. At the same time, scant progress has been made in educating the public, in general, and clients, in particular, about data mining. There are many untruths, half-truths, and downright statistics floating around about how data mining works and how it is used. In this and future articles, I intend to clear up a few of the most pervasive of these misconceptions.

Some misconceptions arise from simple errors in logic. Often, they stem from a lack of familiarity or experience. None are particularly technical problems. All are easily remedied with simple examples and simple explanations. In this article, I will focus on one misconception that I call the "50/50 problem."

An Example of the 50/50 Problem

Recently, I was working with a very bright, energetic client in the biotech industry. Her firm builds imaging equipment and provides services to pharmaceutical companies. The imaging equipment (calling it a complex, microscope-like camera is far too wordy) generated data that she wanted to use to classify chemical compounds as promising or unpromising candidates for drugs. It turns out, in the vast world of chemical compounds, that there are more unpromising drug candidates than promising ones - a lot more. My job was to use data mining techniques to create a classifier (a mathematical formula or a set of rules) that would successfully distinguish promising drug candidates from unpromising ones - using data produced by the imaging equipment.

After some initial work, I presented a classifier to my client. I happily reported that the classifier correctly labeled promising compounds as promising 10% of the time. My client was completely underwhelmed[1]. Her knee-jerk response was, "But you can do 50% just by flipping a coin!"

Actually, a very simple classifier can do much better than 50%. I mentioned earlier that there are many more unpromising compounds than promising ones. In this project, 999 out of every 1000 compounds was unpromising, or 99.9%. A classifier that labels every compound as unpromising is correct 99.9% of the time. Despite its apparently high accuracy, such a classifier is worthless to a pharmaceutical company. Why? Such a classifier would recommend that no compound ever be developed further as a potential drug. Strictly abiding by the classifier, life-saving research would come to an abrupt halt.

The 50/50 Problem in a Nutshell

Is a misconception becoming evident? My client, like many intelligent people, made a simple error in thinking. She made the assumption, because there were two possible outcomes (promising and unpromising), that the outcomes were both 50% likely. This is the "50/50 problem".

My own theory is that many of us are victims of our own education. All of my probability textbooks introduced the subject with discussions about flipping coins. With that as a starting point, perhaps it's no wonder that people make the 50/50 assumption without even thinking about it.

The first step towards a solution is to admit there is a problem. Please repeat after me, "Just because there are two possible outcomes, that does not mean they are equally likely." The second part of a solution is to replace the wrong mental image with the right one. Rather than think about the two outcomes as alternate sides of a coin, think about them as two, clearly unequal, pieces of a pie (see Figure 1). Beneath the crust of the small red piece, picture a filling of delicious fruit. Beneath the large blue piece, picture some mud. Now, consider a blindfolded person plunging a fork into the pie. There are only two possible outcomes, mud or fruit. But the odds of a tasty result are not 50/50, are they?

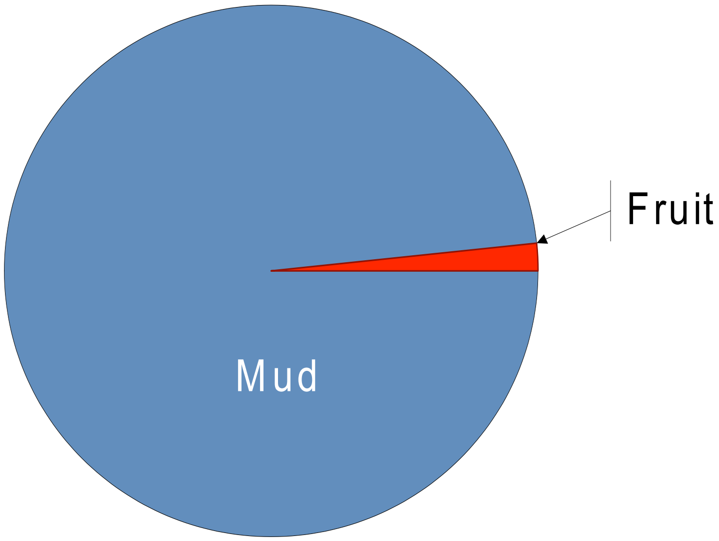


Figure 1 – Two, clearly unequal, pieces of a pie

Data Mining Imperatives

In the drug compound project, the outcomes were not equally likely. One outcome (unpromising) was very common while the other (promising) was rare and significantly more valuable. The rare outcome is the one promising compound out of one thousand candidates that may become a multi-billion-dollar blockbuster drug. In a scenario like this, there are two dominant data mining imperatives:

- To improve the odds of uncovering the rare, valuable outcome
- To make the right kind of mistakes

Improving the Odds

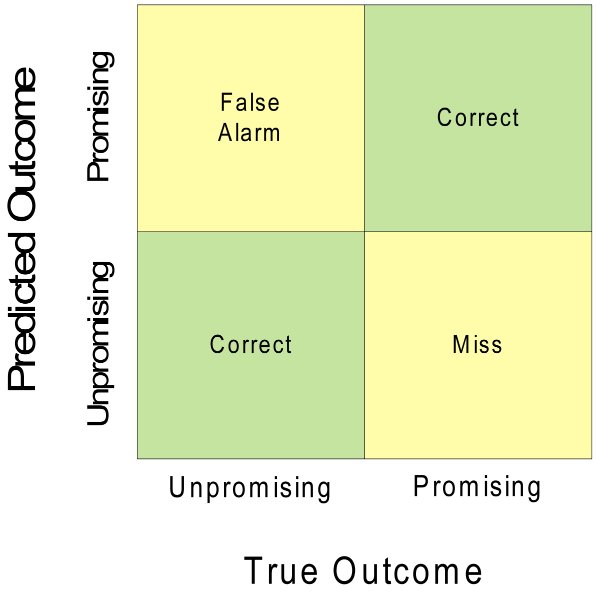
Let's consider each of these imperatives in turn. First, the classifier needed to improve the odds of finding good drug candidates. I mentioned above that a classifier is typically a formula or set of rules. In this case, the drug classifier was a mathematical formula[2]. The inputs to the formula were various measurements of a compound - for instance, the size, shape, or color of biological features in a microscopic image. The output of the formula was a number, often called a score. By design, the higher the score, the more likely the outcome of interest – here, the more likely that the candidate compound would be a promising drug. Provided with a large list of compounds and their associated measurements from the imagining equipment, my classifier produced a list of scores, one for each compound.

Without a classifier, the list of compounds is just a list. A researcher could start with a compound anywhere in the list, synthesize it, test it for promise as a drug candidate, and – 1 time in 1000 - it may turn out to be promising.

With the classifier, the list can be sorted by score, and a researcher can again start at the top, synthesizing and testing compounds. With my classifier, 1 time in 10 a compound near the top of the sorted list will be a promising drug candidate. Allow me to repeat: 1 time in 10. The odds are improved by a factor of 100. Suddenly, the classifier seemed much more powerful to my client.

Making the Right Kind of Mistakes

Having passed the first hurdle by improving the odds, we turn to the second imperative: making the right kind of mistakes. What does that mean – aren't all mistakes created equal? From common experience, though, we know that some errors are worse than others. Taking a misstep in the kitchen is inconsequential. Taking a misstep on a busy city street can have a very serious impact - literally. Let's take a closer look at the mistakes that can be made by the drug candidate classifier.



Predicted Outcome	True Outcome: Unpromising	True Outcome: Promising
Promising	False Alarm	Correct
Unpromising	Correct	Miss

Figure 2 – Predicted versus True Outcomes

See the diagram in Figure 2. One side represents the true outcome, whether a compound is actually unpromising or promising. The other side represents the predicted outcome from the classifier, again, unpromising or promising. (To make this distinction of predicted outcomes, the classifier scores were "cut-off" – compounds with score values above the cut-off value were labeled as promising, while those with values below the cut-off were labeled as unpromising. More about the choice of cut-off shortly).

As shown in green in the figure, there are two ways the classifier can be correct - by predicting promising when the compound truly is promising, and by predicting unpromising when it truly is unpromising. Conversely, as shown in yellow, there are two ways for the classifier to make a mistake: by predicting that a truly unpromising compound is promising (a "false alarm"), or by predicting that a truly promising compound is unpromising (a "miss"). Now, which mistake is worse for a drug maker?

My intuition was that a false alarm was the worse error. After all, I knew that pharmaceutical companies spend many millions of dollars to test and develop a single compound into a drug. Surely, they would want to avoid that expense if they could. While that was true, I learned that a miss was the worse error. Since a promising compound is like a rare gem to the industry, missing any good candidate amounts to missing a potential multi-billion-dollar opportunity. Although costly, some false alarms are acceptable, and even viewed as inevitable - a cost of doing research and business. By setting the cut-off value lower or higher, the classifier can trade misses (billion-dollar missed opportunities) for false alarms (million-dollar dead ends).

Beyond the 50/50 Problem

The 50/50 misconception can fool smart people into thinking that data mining is overhyped, or useless, or both. To counter the misconception, I presented an example of an extremely valuable classifier that was correct just 10% of the time. Conversely, I also described a 99.9% accurate classifier that was useless. Data mining can improve the odds of finding rare, valuable results. Further, with good judgment, data miners can help clients limit the cost of mistakes and missed opportunities. So remember, "just because there are two possible outcomes, that does not mean they are equally likely - or equally valuable."

Tim Graettinger, Ph.D., is the President of Discovery Corps, Inc. (http://www.discoverycorpsinc.com), a Pittsburgh-area company specializing in data mining, visualization, and predictive analytics.

Your comments and questions about this article are welcome. Please contact Tim at (724)-743-3642 or tgraettinger@discoverycorpsinc.com

[1] The reader may also be underwhelmed at this point, but please read on.

[2] In this article, we will ignore the details of how the formula is derived – just know that it can be derived through statistical or other data mining methods.