Tips for Adversarial Analytics

D.B. Skillicorn
School of Computing
Queen's University
Kingston. Canada
skill@cs.queensu.ca

By adversarial analytics, I mean analytic algorithms that model the activities of adversaries as they are captured in data (not the process of learning particular kinds of examples to improve deep learning systems).

These tips contain information that is either not widely known (perhaps outside the intelligence community) or hard to find in a compact format. It will be most useful to those starting out in this field.

- Adversarial analytics are those where the interests of those doing the modelling, and (some of) those being modelled are not aligned. The obvious application domains for adversarial analytics are policing, cybercrime, fraud detection, signals intelligence, counterterrorism, customs and border control, cybersecurity, and anti-money-laundering.
- Adversarial analytics is about generating intelligence, not forensics. Government agencies are used to this difference, but police forces struggle with it. Historically, their approach has been forensic, and they lack the culture, mindset, skills and training to handle intelligence approaches well (intelligence-led policing is more an aspiration than an operational process is most police forces, with some strong exceptions).
- Many more domains are adversarial than first glance would indicate, for example, customer relationship management. Each business wants to estimate their customers' net future value as accurately as possible, while it's in each customer's interest to inflate this, since this will induce better treatment from the business. Differential pricing, where some customers are offered a different price based on their estimated ability to pay, is another adversarial domain. (Businesses based on serving ads purport to act in the interests of both advertisers and potential customers but are, in fact, advertising products that they compute a customer can be persuaded to buy, rather than products that customer might want or need.)
- Adversarial analytics can be used for good – detecting and preventing bad things – but it can also be used for repression.
- The datasets for adversarial analytics tend to come in two forms: those where records associated with adversaries are rare, so that the key problem is to find the needle in the haystack;

and those that are entirely, or mostly, records associated with adversaries and the key problem is to understand adversaries, and perhaps the relationships among them.

- All data analytic work early in a project will discover problems with the data collection process, rather than the traces of adversaries. Even when the data collection path is entirely digital, there will be problems. Often these are because the data passed through either a database or a spreadsheet. Database designers tend not to anticipate all of the possible range of values of data – for example, how long can a phone number be? Spreadsheets tend to apply transformations to data based on what they think the semantics of each field is; worse still they apply these transformations silently. Character set mismatches (ASCII vs unicode) are also often problematic. Different sources tend to use different unicode versions of, for example, single quotes. A new data analytics application should be prepared for this, and the consequent delay in getting actionable results.

- The key difference between adversarial analytics and mainstream analytics is that adversaries are (potentially) aware of the analytics and can take steps in response. These include: avoiding data collection, corrupting the data collection, subverting the data analytics, and social engineering to avoid actions that might result from analytic outputs.

- Adversaries can corrupt the data collection when it is obvious. For example, CCTV cameras collect images and, at first glance, it seems like this ability reduces crime. This turns out to be mostly illusion, since one person in a hoody looks much like any other person in a hoody. Even facial recognition, which does not work as well as the media might suggest, is easily confused by e.g. facial hair. Multiple experiments have shown that the results from surveillance are underwhelming (although there are sometimes advantages *after* an incident).

- Adversaries can subvert the data analytics in two ways. First, they can create records where some of the attributes have the wrong values (wearing a false beard, altering a licence plate). Subverting data analytics is also one of the major motivators for identity theft. Second, they can create records that are unusual and so will tend to be selected for future training. The presence of these unusual records can alter the resulting models.

- Particular analytic techniques are more susceptible to manipulation than others. For prediction, a single carefully chosen record in training data can cause the class boundary of an SVM to move substantially; and, worse still, to move in a predictable way. For clustering, a few records can cause an algorithm such as Expectation-Maximisation to miss outlying records.

- Adversaries can subvert the data analytics by making their records seem as typical or normal as they possibly can. Analytic methods that look for outliers (or even inliers) may fail to find them. For example, distance-based (k-means) or density-based algorithms tend to include close points into the larger group they are similar to.

- Data analytics for adversarial analytics must therefore be re-engineered so that modelling techniques that are resistant to manipulation are preferred. Ensemble techniques, for example, are almost always to be preferred.

- Adversaries can use social engineering to cast doubts on the results of data analytics so that those making decisions based on them will either doubt them or ignore them. For example,

actions can be carried out that are likely to trigger detection as adversaries but are clearly not.

- The ways in which adversaries deal with the prospect of data analytics can become an advantage if the detection process looks for the signs of collection avoidance and data manipulation.

- Deep learning does not help adversarial analytics much. Explainability is less of an issue than in some other domains (because adversarial analytics is about generating intelligence) but deep learning results can contain holes in the results that are undetectable and these are a major concern. These holes tend to be less noticeable than they should be because we, as humans, tend to fill them in without realising it.

- It is always possible to approach adversary detection by matching records to previously known bad records. This is the strategy used by, for example, auditors. However, adversaries are both aware of the conventional signals of bad things, and developing new techniques that do not produce them, so relying on historical patterns of adversaries is playing catchup and can lead to complacency.

- When the task is to pick out the records of adversaries from a large number of other, 'normal' records, the presence of the normal records is key because it provides a backdrop against which the adversary records can be more easily seen. This needs to be communicated better to the public, who are concerned about their data being collected.

- For these needle-in-the-haystack problems, we do not know good algorithms. Techniques such as one-class SVMs seems absurdly sensitive to parameter choice. Outlier detection techniques often fail because the adversary records are either (a) close to the normal records (deliberately), or (b) so-called inliers, that is they lie within the envelope of normal records, a region where few algorithms do well.

- In such datasets, it's usually plausible to make the assumption that frequent equals normal.

- Predictors that rank records by normality are often more useful than classifiers separating normal from adversary. The advantage is that a ranking predictor can be converted into a classifier by deciding on a boundary in the ranking. This boundary can be chosen with knowledge of the ranking and, in particular, the density in the ranking. Thus the boundary does not need to be chosen before the data is examined.

- There's a temptation to attack these rare-adversary records using monolithic approaches. It's often better to use a sequenced approach. During training, the first predictor can be tuned so that it misses no adversary records, even if it predicts many normal records to be adversary. However, those records that it predicts to be normal can be discarded from the second stage. The second predictor is similarly trained so that it misses no adversary records, perhaps still with a large number of normal records predicted to be adversary. Each stage reduces the amount of data that has to be used for training. From a practical point of view, this may enable more sophisticated (and expensive) techniques to be used in later stages. When the predictor chain is deployed, normal records tend to be discarded during early stages; records that make it further down the chain are increasingly likely to be adversary. It becomes feasible to use human judgement as the final stage, even if the initial data is large.

- Applying some randomisation to, for example, choice of boundaries makes it harder for adversaries to judge what behaviour will, with absolute certainty, protect them from discovery. It also makes probing much less useful, since what worked this week in a test may not work next week in an attack. For example, many countries have a $10,000 reporting requirement for cash movements and transactions. Changing this daily would increase the uncertainty for those moving and using cash.

- Insiders who can alter the detection or action processes *within the analytics framework* can be very effective, and are usually considered too far fetched to be a risk.

- Anomaly detection techniques for clustering are not well developed. They tend to fail for adversary records that are close to normal records. Graph-based approaches can be useful because the global structure of a derived graph depends on (almost all of) the local structure.

- Hierarchical clustering can also be effective because small adversary clusters may show up as branches that join the dendrogram late. Its time complexity, however, puts it out of reach of most large-scale analytics.

- Lone adversary records are much harder to detect than groups of adversary records. Fortunately, the needs of adversaries to plan and collaborate mean that their records often have correlations. Thus they tend not to be scattered across the space spanned by the attributes, but (weakly) concentrated. Unfortunately, they tend to be concentrated near the normal records.

- It is hard to distinguish lone adversaries from lone eccentrics, but lone eccentrics tend to be less normal than lone adversaries. Lone adversaries are, at least partly, purpose driven, whereas lone eccentric tend to be less functional.

- It may be difficult for adversaries to know what normality looks like. In an effort to ensure that they seem normal, they may overshoot. This can show up as more exaggerated signals than those of normality, but may also show up as excessive blandness. For example, an adversary may feel the need to have a social network presence, but may post only innocuous content so as not to draw attention.

- "The guilt flee when no man pursueth" – adversaries are more aware of being the subjects of data collection and analysis than ordinary people, and so may react to reminders that this is happening, while ordinary people are either unaware or uncaring. An old pickpockets' trick was to shout "beware of pickpockets" causing everyone to touch the pockets where they were keeping their valuables.

- Because adversaries want to conceal and manipulate their presence in data, the best places to look for them are in data where manipulation is difficult: natural language, and social networks. In general, any area in which the data collected is the result of largely unconscious processes is likely to be fruitful.

- *Purpose tremor* is the result of trying to do something consciously that is normally done unconsciously. For example, most of us are able to insert a key smoothly into a lock (quite a targeting feat) but if we stop and think about it our performance gets worse, not better. Humans are poor at doing the unnatural in a natural way. Amateur actors are those who

are consciously acting, and we see through the character they are portraying. In both cases, conscious thought creates a detectable signal.

- Natural language leaks aspects of the speaker/writer's mental state. This is the basis for authorship detection, but many other properties also leave detectable traces in natural language. These include: deception, age, gender, personality, mood, emotions (but these are very transitory), attitudes to other groups and to products; and many others are being investigated. For example, there is some hope of building models of intent and action.

- Because much of language generation is subconscious, it is hard for adversaries to avoid revealing these properties. Attempts to do so may create countervailing signals of stiltedness or blandness.

- Social networks are difficult to manipulate because the creation and use of a link is (apparently) a local and independent process. However, the aggregate global effect of these local processes create large-scale regularities. For example, an adversary may want to appear normal in a social media setting. How many 'friends' are normal? And what kind of 'friends' should they be? Questions like this are hard to answer if the goal is to seem normal, because ordinary people do not ask them - they choose who to link to in a local, direct way.

- The social networks of adversaries mix with conventional social networks in subtle ways (and the data is hard to collect). Not surprisingly, the risk of being shot if you're a criminal is many times that of the baseline of a so-called stray bullet shooting – but the risk to someone in the immediate social network of a criminal is also much higher than the baseline.

- In datasets consisting entirely of adversaries, conventional data analytic techniques can be used, but there are typically patterns that would not be present in datasets of normality. This is partly because adversaries use tradecraft as a matter of routine. Learning this tradecraft is, of course, one of the reasons to apply data analytics to such datasets.

- Data about incidents of adversary action can provide some insight into their strategy and tactics. At present, simple statistics are often the only analysis.

- A special kind of incident data is that where geographical location is captured. Since movement through space is not trivial, examining positional information can be especially revealing.

- The social network structure of adversaries is useful to understand their group dynamics and command-and-control. It is well known that the social networks of criminals contain many fewer triangles than conventional social networks, presumably because connections are power and they are reluctant to give this away. Co-offender networks of criminals have been studied extensively to understand how criminal planning leads to personnel deployed in crimes. Copresence networks, which connect those at the same incident (whether criminal or not) extend this idea to discover patterns involving bystanders (who are often not involved accidently) and victims.

- Those who must control adversary groups typically have to communicate with one another much more often than foot soldiers. Their region of the social network therefore tends to contain many more triangles than normal (for such a network). This can be used to identify them, using an approach called Simmelian backbones.

- Edge prediction in such networks can find relationships that are deliberately concealed, for example, leaders who do not communicate electronically. It can also identify aliases, that is nodes that purport to be different people but are actually the same.
- Natural language created by adversaries is useful because it can be used to reverse engineer what is on their minds. This can reveal underlying worldviews, intent, and (via authorship) changes in personnel and responsibilities.
- Natural language is also the way in which many adversary groups recruit and influence. Observing how they go about it provides an assessment of competence, and also the underlying framing they use to position themselves in the world.

What have I missed or misstated? Comments, please, to skill@cs.queensu.ca.

David Skillicorn is a Professor in the School of Computing at Queen's University. He has applied data analytics to adversarial settings since the second Bali Bombing. As well as research in the academic community, he collaborates with police forces and government departments in multiple countries. He has published many articles but also the books *Knowledge Discovery for Counterterrorism and Law Enforcement* (CRC Press, 2009) and *Understanding High-Dimensional Spaces* (Springer, 2012).