Data Mining and Privacy...again

by Dean Abbott on 01/05/2010 00:05 0 comments , 125 views
Topics: Data Mining
Tags: dhs

A google search tonight on "data mining" referred to the latest DHS Privacy Office 2009 Data Mining Report to Congress. I'm always nervous when I see "data mining" in titles like this, especially when linked to privacy because of the misconceptions about what data mining is and does. I have long contented that data mining only does what humans would do manually if they had enough time to do it. The concerns that most privacy advocates really are complaining about is the *data that one has available* to make the inferences from, albeit more efficiently with data mining.

What I like about this article are the common-sense comments made. Data mining on extremely rare events (such as terrorist attacks) is very difficult because there are not enough examples of the patterns to have high confidence that the predictions are not by chance. Or as it is stated in the article:

> Security expert Bruce Schneier explains well. When searching for a needle in a haystack, adding more "hay" does not good at all. Computers and data mining are useful only if they are looking for something relatively common compared to the database searched. For instance, out of 900 million credit card in the US, about 1% are stolen or fraudulently used every year. One in a hundred is certainly the exception rather than the rule, but it is a common enough occurrence to be worth data mining for. By contrast, the 9-11 hijackers were a 19-man needle in a 300 million person haystack, beyond the ken of even the finest super computer to seek out. Even an extremely low rate of false alarms will swamp the system.

Now this is true for the most commonly used data mining techniques (predictive models like decision trees, regression, neural nets, SVM). However, there are other techniques that are used to find links between interesting entities that are extremely unlikely to occur by chance. This isn't foolproof, of course, but while there will be lots of false alarms, they can still be useful. Again from the enlightened layperson:

> An NSA data miner acknowledged, "Frankly, we'll probably be wrong 99 percent of the time . . . but 1 percent is far better than 1 in 100 million times if you were just guessing at random."

It's not as if this were a new topic. From the Cato Institute, this article describes the same phenomenon, and links to a Jeff Jonas presentation that describes how good investigation would have linked the 9/11 terrorists (rather than using data mining). Fair enough, but analytic techniques are still valuable in removing the chaff--those individuals or events that very uninteresting. In fact, I have found this to be a very useful approach to handling difficult problems.

---

http://

URL address starting with http://Please type in the digits from the image

Content

Save Cancel

Alert me via email when new comments are added to this post

Bookmarks and share

- Favorites
- Print
- Digg
- Delicious
- Google
- Live
- Twitter
- Facebook
- StumbleUpon
- More

Bookmarks and share

- AIM
- Ask
- Bebo
- Blogmarks
- Buzz
- Delicious
- Digg
- Facebook
- Favorites
- FriendFeed
- Google
- LinkedIn
- Live
- Mixx
- Multiply
- myAOL
- MySpace
- Netvibes
- Newsvine
- Print
- Reddit
- Slashdot
- Spurl
- StumbleUpon
- Tailrank
- Technorati
- Twitter
- Y!Bookmarks