Data Mining in Counterterrorism

David Banks
ISDS
Duke University

1. Context

Many entities are exploring the use of data mining tools for help in counterterrorism.

- TIA (DARPA)
- LexisNexis (airline screening)
- Global Information Group (in the Bahamas)
- Terrorist Threat Integration Center (federal)

These efforts raise legal, political, and technical issues.

From the technical standpoint, there are at least three kinds of use that people have in mind:

- Forensic search
- Signature detection, for various prespecified signals
- Prospective data mining for anomalies

2. Forensic Data Mining

This is the least problematic (socially or legally) and the easiest. It is widely, if a bit inefficiently, used by the police:

- The DC snipers
- Background checks for people who purchase guns
- Post-arrest investigation to build cases (e.g., CSI stuff).

Most forensic data mining problems involve finding matches between known information and large, often decentralized, data.

Examples include record linkage and biometric identification.

Also, one sometimes uses multiple search algorithms, and must combine the signal from each.

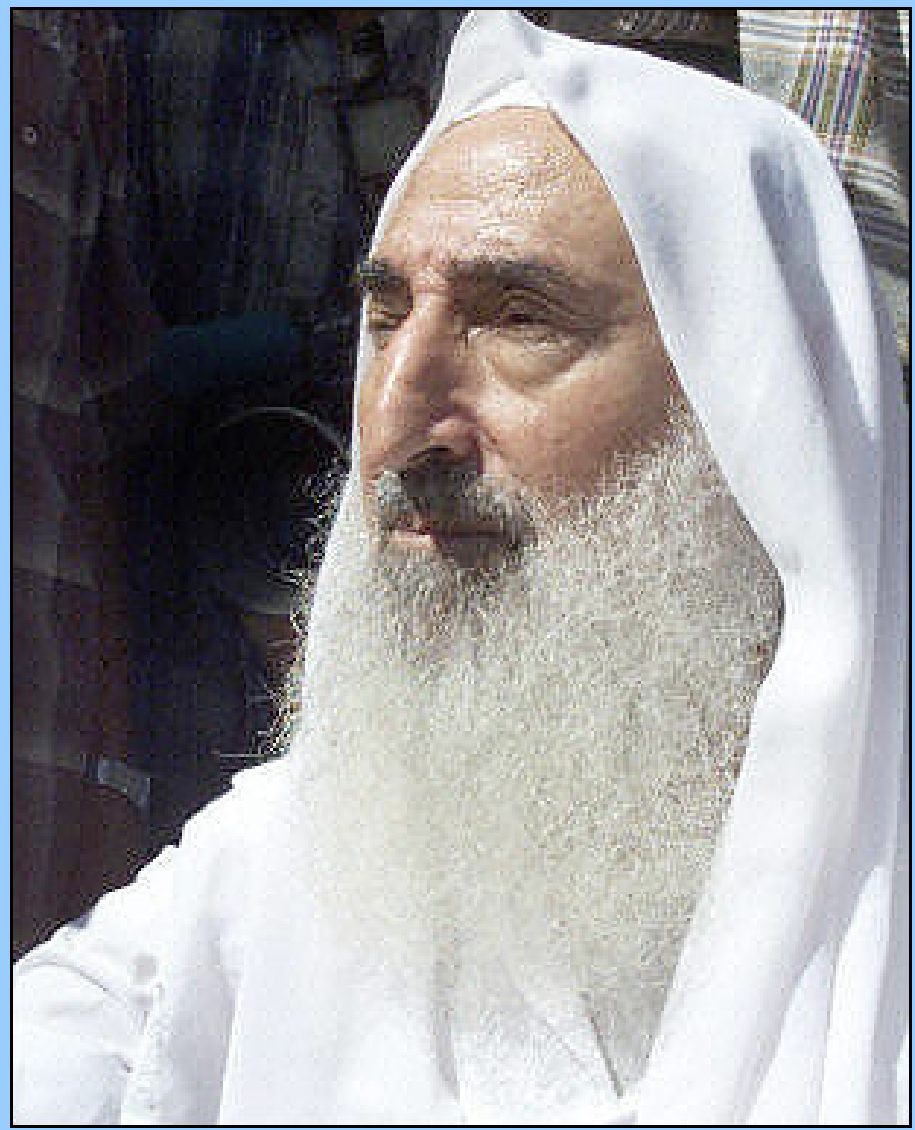
2.1 Record Matching

Biometric identification tries to rapidly and reliably match physical features. This includes fingerprint matching, retinal scans, and gait analysis.

But the ultimate goal is to use AI to match photos to suspect lists.

DARPA and NIST have a joint project to do automatic photo recognition. The project uses a testbed of images called FERRET. (See Rukhin, 2004, *Chance.*)

The main problem is that the feature extraction systems do not work very well. There are high levels of false positives and false negatives. Shadow and angles are hard.



A data mining strategy I like is:

- Have humans assign a sample of photos impressionistic distances;
- Use multidimensional scaling to embed these photos in a relatively low-dimensional space;
- Use data mining to extract the features of the photo that best predict the low-dimensional metric used by MDS
- Use constrained nonparametric regression to fit the metric.

Looking for matches in large and decentralized databases is much like the classical record-linkage problem (Fellegi and Sunter, 1969).

This methodology has become key in many data mining applications:

- Google search
- Census data quality
- TREC studies

The original method tries to formally match partial or noisy names and addresses.

Text matching builds on linkage, with complex rules for stemming words and using co-present text as covariates.

Winkler (2002) is applying modern data mining to linkage (Bayes nets, SVMs, imputation, etc.)

2.2 Combining Algorithms

A common strategy in looking for matches in a database is to use an algorithm to assign a rank or score to the match between the target (say a photo or a fingerprint) and each record in the database.

A human would then assess the top-ranked matches by hand.

One strategy to improve matching is to use boosting (Freund and Schapire, 1996). This approach successively modifies a weak classifier to produce a much better classification algorithm.

In match finding the classification problem is a little non-standard, but the ideas can be adapted.

The boosting algorithm is:

1. Initialize obs. weights $w_i = 1/n$
2. Do m=1, …, M times:

a) Fit $G_m(x)$ with weights w_i

b) Get $e_m = \Sigma\, w_i\, I(\text{error on } x_i) / \Sigma\, w_i$

c) Compute $\alpha_m = \log[(1 - e_m)/em]$

d) Use $w_i \exp[\alpha_m\, I(\text{error on } x_i)]$ in place of w_i

3. Use $G(x) = \text{sign}[\Sigma\, \alpha_m\, G_m(x)]$

The boosting algorithm can be viewed as a weighted sum of M related algorithms.

Similarly, one can use weighted combinations of unrelated algorithms; this also improves accuracy.

But there are hard issues on how to weight or combine.

Rukhin (2004) describes issues in combining algorithms for biometric identification.

Using scan statistics, copulas, and partial rankings, he finds a general procedure to combine algorithms to provably improve accuracy for top-rank matches.

One might also try a partitioning approach.

3. Signature Search

Tony Tether at DARPA said that IAO was being used for detecting preset profiles---e.g., Arabic men studying in flight schools.

This is different from forensic search, because the analysts get to pick what they want to find.

DARPA had the TIA/IAO program, the NSA and the DSA have(?) programs, and so forth. The intention is to combine public, federal, and private databases and use statistical techniques to find prespecified risk behaviors.

Contractors have commercialized some programs of this kind, or even taken them off-shore.

The tools used are similar to those needed for forensic datamining. The main distinction is that instead of searching for a match, one has to search for records in a neighborhood of the signature.

This is because terrorists will try to smudge the signal. This could entail false addresses and names, indirect travel patterns, etc.

The tools used are similar to those needed for forensic datamining. The main distinction is that instead of searching for a match, one has to search for records in a neighborhood of the signature.

This is because terrorists will try to smudge the signal. This could entail false addresses and names, indirect travel patterns, etc.

In signature detection problems one probably needs a human in the loop.

One model is the type of success that El Al screeners have had.

Data mining can do a lot of preliminary sieving, but humans have domain knowledge that is hard to embed in an AI system.

4. Prospective Mining

This is the Holy Grail in data mining for counterterrorism. One wants to be able to automatically “connect the dots”.

This is extremely hard and perhaps infeasible. The most mature application is in cybersecurity.

Cybersecurity rides two horses:

- Signature detection (CERT, McAfee, spam blockers)
- Anomaly detection.

Anomaly detection tries to identify hack attacks that have not been previously seen. And this has to be done automatically, in order to be sufficiently fast.

DARPA and others have been doing research in anomaly detection for years. The big problem is the false alarm rate. In a high-dimensional space, everything looks funky.

False alarm rates have been a thorn in the side of other federal security programs. The recent NRC study on the use of polygraph data found the error rate was just too high.

Suppose that:

- the false alarm count is f, and the average cost of a false alarm is c.
- the number of missed alarms is m, and the average cost of a missed alarm is d.
- the cost of the program is p.

Then the system should be used if

$$f*c + p < m*d$$

This decision-theoretic formulation is much too simplistic. One can rank the alarms, do cheap preliminary tests, and use human judgment.

Nonetheless, this is the right way to assess a system. Any agency that has such a system can easily check whether their program is cost-effective.

5. Conclusions

Data mining has a major role in modern counterterrorism. It is a key methodology in:

- Syndromic surveillance and emergent threats
- Record linkage for background checks
- Cybersecurity

We want to extend the reach of data mining to include:

- Biometric identification
- Automatic dot connection
- Social network discovery
- Prospective classification

But these are hard problems, and some may not be possible. The U.S. government needs to be realistic about what is possible.