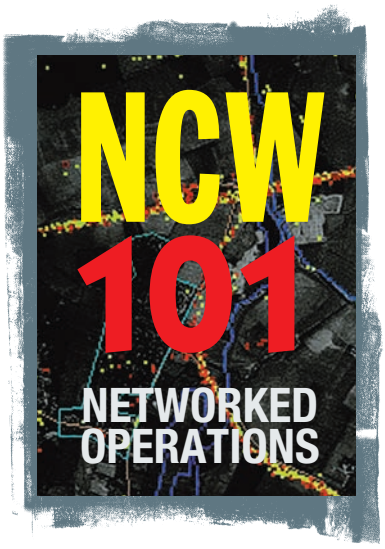

Data Mining and Knowledge Discovery techniques

NCW 101 Part 17

Dr Carlo Kopp

The advent of low cost computing power and fast storage capacity has opened up a range of techniques that have historically been expensive to implement and deploy as operational capabilities. Data Mining and Knowledge Discovery techniques fit this pattern and are being used with increasing frequency in the commercial sector, as market analysts invest increasing time and effort to identify new product niches.

In a heavily networked environment where adequate sensor capability and proper integration exist, inevitably the problem that will arise is making sense of abundant raw data, extracting from that data what is actual information, and validating that information so as to produce knowledge and understanding.

Automated software mechanisms are inevitably required to assist the human analyst and accelerate the process of converting raw data into a situational picture.

One of the appealing misconceptions which has been widely propagated by non technical advocates of military networking is that this type of effort can be sufficiently automated, so that a human decision-maker can see an almost immediate situational picture and be able to action it immediately. This is something that for the forseeable future will remain in the domain of science fiction movies. Despite decades of research effort Artificial Intelligence (AI) technology is still not capable of emulating human cognitive processes, to the extent that we can replace imagery analysts and intelligence officers with a few CD-ROM's worth of unusually clever software.

Computers are very good at doing large volumes of repetitive work, including work that otherwise requires a lot of computational effort. Computers, in this era of ridiculously cheap disk storage capacity are also very good at searching vast volumes of data, be it in properly structured databases or in other formats.

The techniques known today as Data Mining and Knowledge Discovery evolved over decades from a range of well established techniques in statistical processing of digital data.

Unfortunately, the commercial nature of the computer industry has also seen widespread hijacking of terminology to make pedestrian software products appear to be more than what they really are, so as to maximise sales. There are for instance far too many word processing software packages in today's market, being sold as 'desktop publishing suites'. This is no different than the defence industry, where the mediocrity of many products is well hidden behind a facade of flowery marketing labels.

Data mining in particular has become a victim of this effect, to the extent that a good number of very conventional techniques for searching large sets of data are being marketed as 'data mining'. Novices in this area need to therefore tread very carefully.

What is Data Mining and Knowledge Discovery?

Data Mining is defined as "the process of extracting trends or patterns from data" (Wright, 1998), or in the words of one practitioner, "the process of hypothesis generation". As such, data mining involves having suitable software 'crunch' its way through a large set of relevant data to identify and isolate relationships that exist but may not be obvious from an immediate examination of the dataset. Whether what the data mining process discovers is useful, relevant or exploitable depends on what is discovered. The label of 'hypothesis generation' is valuable here, insofar as it emphasises the fact that data mining identifies possibilities which might be exploited, allowing a more focused and directed analysis effort to be applied to determine whether the proposition is indeed true, and how specific or general it might be. Simply fishing the obvious out of a large mass of data is not data mining, it is simply data retrieval, despite what the marketeers might claim.

Knowledge Discovery is a technique that can be applied to the results of data mining, to make sense of them. Frawley et al define knowledge

discovery as "the non-trivial extraction of implicit, unknown, and potentially useful information from data", and usually this is assumed to involve the use of artificial intelligence techniques.

Given how closely coupled data mining and knowledge discovery are, more than often we see the terms used interchangeably, or either term used to describe both techniques. This is not an uncommon problem in emerging or immature disciplines, and parallels other 'nomenclature aliasing' cases in the computer industry.

In a survey paper Wright presents six factors observable in knowledge discovery techniques:

- All approaches deal with large amounts of data
- Efficiency is required due to volume of data
- Accuracy is an essential element
- All require the use of a high-level language
- All approaches use some form of automated learning
- All produce some interesting results.

"Large amounts of data are required to provide sufficient information to derive additional knowledge. Since large amounts of data are required, processing efficiency is essential. Accuracy is required to assure that discovered knowledge is valid. The results should be presented in a manner that is understandable by humans. One of the major premises of KDD is that the knowledge is discovered using intelligent learning techniques that sift through the data in an automated process. For this technique to be considered useful in terms of knowledge discovery the discovered knowledge must be interesting; that is, it must have potential value to the user."

That these techniques work, and often work well, is well established in the literature. Some very good case studies exist in the commercial domain.

Yale University academic Ian Ayres' 2007 book 'Super Crunchers' is a popular study of what can be achieved using these techniques, and includes a good number of examples of what has been achieved in practice using such techniques. It also underscores the intensive conflicts that have arisen between practitioners of digital prediction and classical experience driven human analysts objecting to these intrusions into their hallowed turf.

One good case study is that of US economist Orley Ashenfelter, who used statistical techniques to analyse the quality of French Bordeaux wines. Specifically he sought to relate auction prices to specific local annual weather conditions, in particular rainfall and summer temperatures. His findings were that hot and dry years produced the wines most valued by buyers. Ashenfelter's work and analytical methodology resulted in a deluge of hostile invective from established wine tasting experts and writers. Ayres puts this down to fear of losing a lucrative monopoly, and the reality that a better informed market is more difficult to manipulate on pricing.

Another interesting case study that Ayres discusses is that of US baseball analyst William James, who applied statistical methods to the prediction of which players would be most successful in the game, challenging the traditional approach followed by talent scouts who observed prospective new players and made predictions on the basis of observed technique with the bat. Not unlike AFL, rugby or cricket in Australia, baseball is a major domestic industry in the US, and there are lucrative earnings to be made from locating and recruiting the right talent for teams. James' statistically driven approach to correlating early performance to mature performance in players resulted very quickly in a barrage of criticism, not unlike that seen in the wine industry. As an interesting aside, this methodology has since acquired the name 'sabermetrics' and was popularised in at least two episodes of the 'Numb3rs' television series.

Ayres details another nice case study of a research project he conducted jointly with Steven Levitt, to determine the impact of LoJack radio beacons on car theft statistics. The LoJack system is used to physically locate stolen motor vehicles, permitting law enforcement to find the vehicle and more than often the thief as well. What Ayres and Lovitt determined was that LoJack produced a strong deterrent effect and as a result, a strong reduction in car theft in areas where it was used. What they also determined was that car insurers provided discounts to LoJack users which were much smaller than discounts which would reflect the savings they earned from reduced numbers of claims, the consequence of LoJack use in that geographical area.

Yet another case study detailed by Ayres are dating services. For instance, the eHarmony dating service, which rather than matching prospective partners on the basis of their stated preferences, uses statistical analysis to match prospective partners, based on a 29 parameter model derived from 5,000 successful marriages. Its competitors such as Perfectmatch use different models, such as the Jungian Meyers-Briggs personality typing technique to parametrise individuals entered into their database. It is worth observing that while the process of matching partners may amount to little more than data retrieval using some complex set of rules, the process of determining what these rules need to be involves often complex knowledge discovery and mining techniques.

That the use of statistical methods to divine useful or interesting information or knowledge from large masses of data has been a commercial success where applied intelligently is beyond dispute. Ayres points out that the increasing use of such techniques on various web search engines to suggest other links of potential interest (eg products) is a demonstration of how any technology with the potential to increase profits will be adopted.

Given the established usage of such techniques in the commercial sector, what value do they really have in the military domain?

Military Applications of Data Mining and Knowledge Discovery

There have been claims that data mining and knowledge discovery techniques have been used successfully in counter-terrorism intelligence analysis, but little has surfaced to support these claims. The idea is that by analysing the characteristics and profiles of known terrorists, it should be feasible to predict who in a sample of population might also be a terrorist.

This is actually a good example of potential pitfalls in the application of such analytical techniques to practical problems, as this type of profiling generates hypotheses, for which there may be good substantiation, of the ilk of 'Joe Bloggs is a possible terrorist'. The risk is that overly zealous law enforcement personnel, again highly motivated for good reasons, over-react when the individual despite the profile is not a terrorist. There is enough evidence in the media, albeit sensationalised, to suggest this is a real risk.

A hypothesis is not a validated and proven fact, it is essentially a proposition, which remains to be proven. It might be correct, but it also might not be. The analytical method has thus divined a possible outcome, but only careful investigation can prove whether the possibility is a probability, or indeed a certainty with a probability of one.

Some valid criticism has been directed over the years at NCW zealots for precisely this reason. The ISR output gathered by a networked system will provide a large number of known targets, but possibly an even larger number of uncertain or possible targets. How does one reliably divine which of the uncertain or possible targets are real ones? And what are the consequences of putting a smart munition into an uncertain target, especially if significant collateral damage or unintended damage occurs? The sorry history of 'friendly fire' incidents and erroneous targeting over the last century of modern warfare speaks for itself. Certainty is not a cheap commodity, especially in the environment of modern high tempo combat where targets are fleeting, and firing opportunities may be scarce. The temptation to take a shot before the opportunity is lost can be overwhelming, as history proves again and again.

That data mining and knowledge discovery techniques promise exceptionally high value to the ISR and intelligence communities cannot be disputed. In particular, the application of such techniques for the identification of possible targets, which satisfy some valid criteria, has the potential to save vast numbers of personnel hours. More than often in recent conflicts the capacity to deliver firepower well exceeded the capacity of the ISR system and its supporting analytical personnel pool to find, validate, classify, prioritise and generate targets for attack.

What networking and abundant ISR does deliver is the volume of raw data necessary to match available firepower delivery capabilities. What networking and abundant ISR do not deliver is the analytical capability to transform the raw ISR output into actionable targeting information, and integrate the potential use of that information with a direct military outcome.

Where data mining and knowledge discovery techniques offer an immediate and high military payoff is in generating rules for automated target recognition and classification. Much of the business of automated target recognition is about correlating some collection of target signatures and attributes against a known set of parameters. As camouflage and stealthing techniques become increasingly available, the quality and volume of gatherable signature data will inevitably decrease. As a result, finding targets with a high level of confidence will become more difficult. Moreover, the popular technique of 'human shielding' or hiding military targets inside civilian structures put a premium on sorting the sheep from the wolves.

The reality is that data mining and knowledge discovery techniques have other less obvious applications with a high military payoff, even if this payoff lies outside the scope of networked combat.

One of the persistent problems the military in Western nations has faced for decades is in recruitment, and specifically in recruiting and streaming recruited personnel to best exploit the available pool of talent. The classical case study is in how to pick as early as possible the recruits who will make good pilots, and of these, who have the talent to be successful as fighter pilots or bomber pilots, since not every candidate pilot has been gifted with the right mix of talents. As one senior RAAF officer involved in fighter pilot training observed to this author some years ago, the ultimate determination of suitability for a fighter squadron posting was often not made until the candidate was flying an F/A-18 Hornet in the Operational Conversion Unit. More than often, promising candidates did not perform to expectations and had to be posted elsewhere. What data mining and knowledge discovery techniques offer here is the means of analysing decades of Service personnel data to produce more accurate profiles, which can in turn be exploited in the recruiting and training process. Significant training costs could be saved if this is performed successfully.

Another area where such techniques offer promising gains in efficiency is in the maintenance of military platforms. Good and analytically based maintenance programs, with the Amberley Ageing Aircraft Program for the F-111 a good example, systematically analyse component failure statistics to identify components with wearout or other failure rate problems. They can then be removed from the fleet by replacement with new or re-engineered and thus more reliable components. This type of analysis is a simple rule-based approach, where the rule is simply the frequency of faults in specific components. What it may not discover is where the failure prone components produce excessive wear and stress in other components, causing them to fail sooner. No differently, basic reliability analyses do not necessarily correlate specific failure frequencies of components with the platform's operating regime. If road transit in a tracked armoured vehicle increases the statistical odds of a specific electronic component failing due to vibration induced damage, it might not be obvious from cursory analysis.

Data mining and knowledge discovery techniques have the potential to yield very high payoffs in the military game, no differently than observed in the commercial sector.

Further Reading:
http://www.randomhouse.com/bantamdell/supercrunchers/
http://www.acm.org/crossroads/xrds5-2/kdd.html