Data Mining Case Study: Modeling the Behavior of Offenders Who Commit Serious Sexual Assaults

Richard Adderley
West Midlands Police
Queens Road Police Station
Birmingham, B6 7ND, UK
Tel +44 121 322 6010
r.adderley@west-midlands.police.uk

Peter B. Musgrove
University of Wolverhampton
35-49 Lichfield St.
Wolverhampton, WV1 1EL, UK
Tel: +44 1902 321851
P.B.Musgrove@wlv.ac.uk

ABSTRACT

This paper looks at the use of a Self Organizing Map (SOM), to link of records of crimes of serious sexual attacks. Once linked a profile can be derived of the offender(s) responsible.

The data was drawn from the major crimes database at the National Crime Faculty of the National Police Staff College Bramshill UK. The data was encoded from text by a small team of specialists working to a well-defined protocol. The encoded data was analyzed using SOMs. Two exercises were conducted. These resulted in the linking of several offences in to clusters each of which were sufficiently similar to have possibly been committed by the same offender(s). A number of clusters were used to form profiles of offenders. Some of these profiles were confirmed by independent analysts as either belonging to known offenders or appeared sufficiently interesting to warrant further investigation.

The prototype was developed over 10 weeks. This contrasts with an in-house study using a conventional approach, which took 2 years to reach similar results. As a consequence of this study the NCF intends to pursue an in-depth follow up study.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications - *data mining*

General Terms

Experimentation, Verification

Keywords

Knowledge Discovery, Data Mining, Crime Pattern Analysis, Offender Behavior, Self Organizing Map.

KDD 01 San Francisco CA USA
Copyright ACM 2001 1-58113-391-x /01/08...$5.00

1. INTRODUCTION

Data mining is now a proven technology in many industrial sectors. The challenge now is to extend the range of applications to which data mining is used in order to both spread benefits and investigate any domain specific problems that require enhancements to current data mining practices.

Police forces across the developed world have attempted to apply advanced computing technologies to tackling crime[2]. However, comparatively little use has been made of data mining techniques in the analyzing and modeling the behavioral patterns which occur in the commission of a crime.

This paper looks at applying data mining techniques to the task of linking crimes of a serious sexual nature. The challenge being to decide which of the separate offences can be linked as being possibly committed by the same offender(s). The intent being to link offences based on coded data (see section 4) and to subsequently produce a profile of the offender(s) that describes the linked theme.

This work draws on an earlier study applied to linking crimes of burglaries due to the offender(s) passing themselves off as a bogus official in order to gain access to a dwelling with the intention of committing theft[1]. The current study was conducted on a larger scale and provided with more resources, which enabled the system to be supplied with cleaner data.

The commercial data-mining package SPSS Clementine was used in order to speed development and facilitate experimentation within a Cross Industry Platform for Data Mining (CRISP-DM) [6] methodology. This enabled the prototype system described in this paper to be developed in ten weeks. This contrasts with an in-house study using conventional techniques, which lasted two years and produced similar results.

In this paper the self-organizing map[9] technique is used to analyze sexual assaults and rape offences held in a ViCLASS[1] relational database within the National Crime Faculty (NCF) at Bramshill, the National Police Staff College. The stages of data selection, coding and cleaning are described together with the interpretation of the results.

[1] Violent Crime Linkage Analysis System

2. TASK UNDERSTANDING

When a specified offence occurs within the United Kingdom (UK) the Force in which the offence occurred has the remit to forward full details to the NCF for subsequent entry onto the ViCLASS system. A specified offence includes a sexually motivated murder, rape where the offender is a stranger or only has limited knowledge of the victim, abduction for sexual purposes and serious indecent assaults. ViCLASS is a relational database developed in 1991 by the Royal Canadian Mounted Police comprising 53 tables not all of which are used in the UK. The system not only stores hard factual information relating to the crime but the offender's behavior is also encoded. Trained analysts examine a 165 question input document and extract behavioral information from narrative text such as the offender's speech and physical actions immediately prior to and during the commission of the crime.

On receipt of the document an Analyst Assistant uses a quality control document for guidance to ensure a consistent approach to data interpretation. It is the role of the NCF Analysts to examine each new case, the index case, with a view to identifying similarities with existing offences within the system. If such links are made it can identify that the index case is part of an emerging series of crimes committed by the same offender(s) who may or may not be known. If a specific series cannot be identified, the analysis may still reveal similarities with other crimes that will assist the Senior Investigating Officer in the investigation of the index crime. Current Police crime recording systems do not transcend individual Police Force boundaries, therefore, those crimes that occur in different Force areas are more difficult to detect. It is within the NCF remit to provide additional assistance in these circumstances.

3. DATA UNDERSTANDING

A copy of the database used in this study contained 2370 recorded sexual offences that occurred throughout England, Scotland, Wales and Northern Ireland between March 1998 and June 2000 and were referred to NCF for analysis.

Table 1 shows the eight specific crime categories and the numbers of offences associated with each. The categories are not mutually exclusive, an example being an offender who commits a burglary with a view to sexually assaulting the victim.

Table 1 - Classification of Sexual Crimes

Offence Category	Date Rape	Burglary	Sexual Assault	Multiple Offenders	Abduction	Weapon	Aggravated Assault	Other	**Total**	Offence Total
Total	22	138	1786	223	230	306	339	266	**3310**	2370

There are 1015 known offenders (convicted, charged or suspected) 90 of whom are believed to have committed 2 or more offences.

Table 2 - No. of Crimes Attributed to Known Offenders

No of Offences	No of Offenders
18	1
13	1
10	1
9	1
7	1
6	1
5	7
4	8
3	9
2	60

Table 2 shows the known prolific nature of offences committed by these 90 offenders. It is possible, however, that some of the undetected crimes in the database may be attributed to these known offenders. Also the offenders of the crimes reported in the database may have committed other similar crimes that have not been reported to the NCF for analysis.

Offences committed by persons unknown to the victim are particularly difficult to detect. However, within a series the offender's behavior often has consistencies across the crime set [3,11]. There is a tendency for the levels of violence to escalate across time[8, 10].

In attempting to model offender behavior by linking crimes it is necessary to understand certain fundamental limitations:-

- The data set is not complete. Although all forces have the remit to forward specified cases to the NCF it is apparent that this does not always occur.
- It is possible that unsolved crimes held within the database may be attributed to one of the known offenders.
- Although the same team of people has input the crimes there are discrepancies in the encoding process, which are discussed below.
- Additional information that could be used to identify similarities between the crimes is held in free text 'memo' fields within the database, they are not currently used in the modeling process.
- The series identified by NCF Analysts for which no offender has been charged, suspected or convicted are assumed to have been committed by the same person(s) for purpose of verification of the models.
- It has to be assumed that the known offenders have actually committed the crimes that have been attributed to them.

4. DATA PREPARATION

4.1 Ambiguous Data

The quality of the results of the mining process are directly proportional to the quality of the data. With a small number of persons responsible for encoding and entering the data it was assumed that the quality would be high. However, there were some discrepancies within the subsequent encoding. Table 3 illustrates an example of confusing encoding of free text information. In this example the variable being encoded relates to whether the victim was specifically targeted as an individual (not just targeted due to the type of person they were). It is clear that both the Yes and No contain the same information and all should have been encoded as No.

Table 3 - Six Data Encoding Examples

Specifically Targeted = Yes	Specifically Targeted = No
The intention of the group involved in this offence was to pick up a prostitute - so to that extent she was targeted.	Required prostitute but the individual was not targeted
In that she was a prostitute however it need not have been specifically her.	The offender did not target that particular prostitute.
As being a vulnerable female	Only in as much as she was a single, young, vulnerable female.

4.2 Missing Data

It is not uncommon for the encoded data to have fields that contain unknown or missing values. There are a variety of legitimate reasons why this can happen. In this specific task one such occurrence might be due to the victim not recalling certain facts due to the trauma associated with the crime. How should they be treated? Are those fields essential to the mining process? There are a number of methods [12] for treating records that contain missing values: -

1. Omit the incorrect field(s)
2. Omit the entire record that contains the incorrect field(s)
3. Automatically enter/correct the data with default values e.g. select the mean from the range
4. Derive a model to enter/correct the data
5. Replace all values with a global constant

Within this study both missing and unknown data have been set to zero when used in dichotomous variables.

4.3 Data Encoding

Data was encoded as either 1 of n or binary dichotomous variables. Categorical data was encoded by sets of mutually exclusive binary variables. An example being the offender's build which was encoded as one of: Unknown, Thin/Skinny/Slim, Medium, Heavy/Stock/Fat being encoded as four binary variables with one variable set to one and the remainder to zero.

Three of the data fields each contain a large number of options, some of which appear to be "close" in meaning :-

1. The approach – the type of behavior that the offender exhibited at the beginning of the crime
2. The precautions that the offender took during the commission of the crimes
3. The verbal themes; speech used during the crime.

There are 29 options for the approach classification, 22 for precautions and 28 for the verbal themes options. Encompassing research [4,7] the approach options have been reduced to three mutually exclusive dichotomous variables. In conjunction with extensive discussions with the analysts the precautions options have been reduced to four mutually exclusive dichotomous variables and the verbal themes reduced to seven fuzzy dichotomous variables.

4.4 Variable Selection

It is always difficult to ascertain the correct number and combination of variables that are to be used in the modeling process. Within this paper it is the intention to model offenders' behavior to establish consistency across crimes. To test this, two different sets of paired variables representing particular behavioral traits were used.

4.4.1 Exercise 1

The first modeling exercise used only the approach and verbal themes sets of variables. This combination was selected to examine behavioral traits at the initial offender/victim point of contact and the subsequent dialogue throughout the crime. This resulted in a total of 3 approach ("con", "surprise" and "blitz") and 7 verbal themes variables being used in training the model.

4.4.2 Exercise 2

Research conducted on male offenders who have committed rape offences within the south of England [5] established that they committed their offences close to their home base or similar suitable anchorage point. The second modeling exercise was restricted to a single Police area. The variables of approach and type of precautions taken by the offender, was used as input. This resulted in a total of 3 approach and 4 precaution variables being used in the model.

5. Model Building

A Kohonen [9] self-organizing map was used in the modeling process for both exercises discussed above. A self organizing map (SOM) was selected because it has the ability both to cluster similar records in to the same cell whilst producing a two dimensional topological map showing the relationship of those records to near neighbors. This can be used to form larger clusters by merging neighboring cells [1]. It also aids in determining the relationship between broad categories of crime. In this application this could be useful as crimes that are broadly similar may have been split in to different clusters due to slight variations in offender behavior due to the specific circumstances in which the crime was committed or even due to missing data.

5.1 Exercise 1 Model

Figure 1, illustrates the 2D representation of the resulting modeling process for Exercise 1. The map shows the 2370 crimes on a 20 by 20 grid (400 cells) as points agitated within a cell to show density. Whilst there is a degree of variability in terms of how approach types, con, surprise and blitz, have been distributed they do appear to fall in to three broad regions. Lines have been manually drawn on the graph to show these regions. Within each area there is further structure due to the verbal behavior exhibited. (An alternative automated approach to manually forming these regions would have been to use a Multi-Layered Perceptron to take the output coordinates from the SOM and label the regions appropriately). The clusters enclosed by square shapes contain those crimes from each of the "approach" types randomly selected from within each broad area for independent verification (section 6).

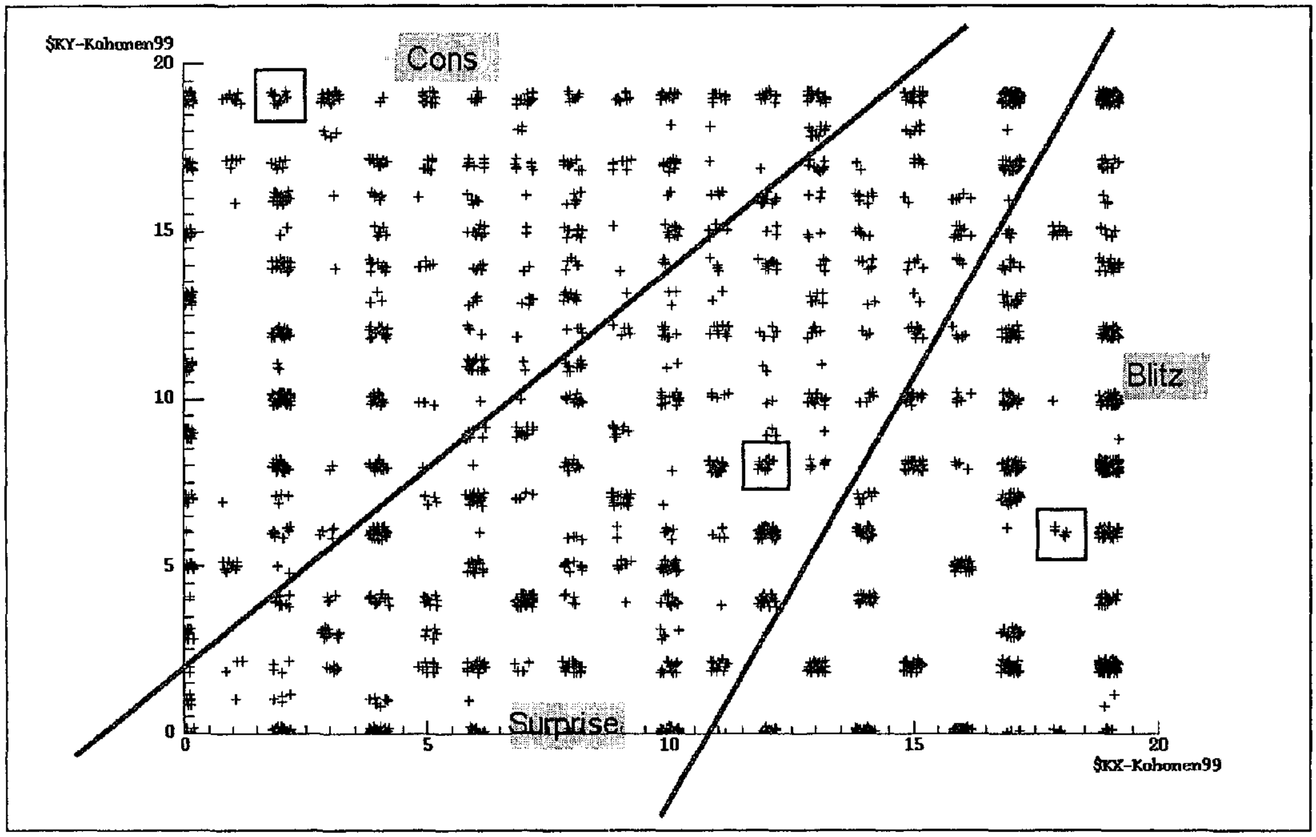

Figure 1 Self-Organizing Map of Approach and Verbal theme offender behavior

5.2 Exercise 2 Model

Figure 2, illustrates the 2D representation of the resulting modeling process for exercise 2 on a 10 by 10 grid (100 cells). Again individual crimes lying within a cell have been agitated to show density.

There are fewer crimes in this model as the geographical region was restricted to encompass a single police area. Again lines have been manually drawn to broadly separate the differing types of approach that the offender used at the point of contact with the victim.

The triangle contains fifty four crimes belonging to a super-cluster that were selected for independent verification. The particularly dense clusters at the top of the SOM have occurred due to missing data rather than a well-defined behavioral pattern. This illustrates the effect missing data can have on the modeling process.

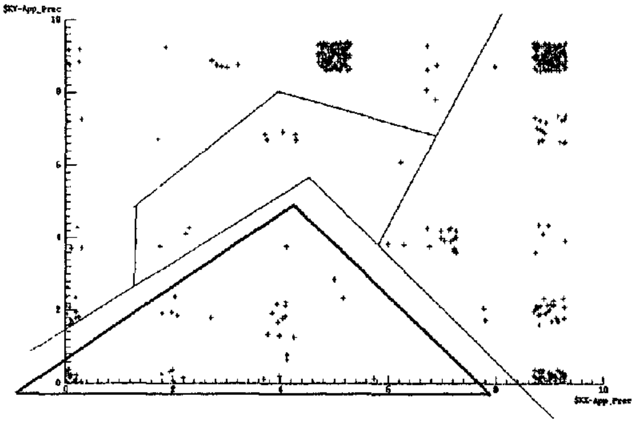

Figure 2 Approach and Precautions Behavior

6. Verification

NCF analysts who took no part in the modeling undertook the verification process but due to their workload they could not examine all crimes within all clusters. Each analyst was only presented with a list of crime identification numbers, with a remit to ascertain whether there were similarities between the crimes, they had no other information. The narrative (discussed in Section 2) was mainly used to ascertain the similarities between the crimes.

6.1 Exercise 1

Three clusters represented by the squares in Figure 1 above were sent for independent verification. In exercise 1 the initial clustering process used the Approach and Verbal themes variables and it was established that there were additional similarities between the crimes contained in each cluster.

The similarities in cluster 1 consisted of the following: -

- 80% of the victims were under the influence of alcohol
- The type of sexual assault was 100% consistent
- Precautions were taken by the offender in 80% of crimes
- Although the offender immediately overpowered the victim on contact, only minor injuries were caused in 80% of the crimes

The similarities in cluster 2 consisted of the following: -

- 50% of the offenders were of the same non-white race
- 53% of the victims were walking in public places at the time of the offence
- A further 33% of victims were asleep at the time of the attack
- There were 2 partial series contained within this cluster
- 4 crimes were part of a known series

The similarities in cluster 3 consisted of the following: -

- The victim was subjected to a number of sexual acts in 100% of the crimes
- 100% of the offenders took precautions
- In 100% of the crimes the offender disrobed himself as well as the victim

Of the three clusters submitted for validation, cluster 2 contained the largest number of offenders and cluster 3 the fewest. It would appear that the number of crimes contained within the clusters indicate the accuracy of the clustering process; the fewer the crimes in the resulting clusters, the more similarities there appear to be.

The three clusters bounded by squares in figure 1 were also examined using extra variables shown in table 4 together with a 'control group' formed by Monte Carlo simulation. The purpose of the control group is to provide a description of a "typical" sexual offence in the belief that variations from typicality could indicate significant behavioral characteristics.

The headings in table 4 refer to the attributes in the database tables, for example 'Questions' refers to whether the offender questioned the victim.

Table 4 - Cluster Comparison

Cluster	Same Sub Approach	Same Sub Verb Theme	Negotiation	Disrobing	Reassurance	Questions	Victim Build	Victim Marital Status	Victim Drug
1	91%	36%	82%	60%	91%	92%	40%	60%	80%
2	75%	100%	75%	25%	50%	50%	24%	100%	25%
3	36%	86%	50%	14%	50%	28%	70%	70%	40%
Control Cluster	19%	38%	73%	30%	71%	59%	44%	58%	27%

Both Approach and Verbal themes variables, described in Data Encoding above, comprise a number of sub variables that, individually, are used in this table for comparison purposes. An example being, 91% of offenders in cluster 1 used the same sub approach type on initial contact with the victim comparing with the control group of 19%. This is, therefore, a significant behavioral trait appertaining to that cluster.

It is important to note that such traits that fall below the typical are also significant in identifying individual offenders e.g. only 28% of offences in cluster 3 try to question their victim during the commission of the offence, comparing with 59% in the control group. This indicates that the majority of offenders question their victim in some way whereas the offenders in cluster 3 do not thereby identifying a particular behavioral trait for that cluster.

6.2 Exercise 2

The crimes belonging to the single approach type that are captured within the triangle in Figure 2 above were passed to an analyst for verification purposes. Individual clusters were not identified due to the group of 54 crimes being considered a 'Super Cluster'.

Within the triangle a complete crime series of four and five partial series were identified and the following similarities were found: -

- 92% of offences were committed by person(s) unknown to the victim
- 59% of offenders had the same motive
- Full intercourse took place in 51% of offences
- 46% of victims were fondled by the offender
- 41% of offenders committed burglary to commit the offence
- 38% of offenders were concerned about their own safety
- 35% of offenders used the attack to satisfy their ego/pleasure
- A weapon was visible in 32% of offences

None of the above was used in the modeling process.

7. DISCUSSION

Given the limitations identified in Section 3 the results are encouraging. In both exercises, based on the results of the verification process, there is some evidence that the models identify consistency in offender behavior. The analysts established that crimes in individual clusters exhibited strong similarities, with adjacent clusters that are based on a variable theme having similar traits as illustrated in Figure 2.

An analyst currently compares the index case against the remainder of the database by selecting one or more variables from the screen, which is then translated into a SQL query. This process relies on the analyst's skill and intuition often resulting in a time consuming process of multiple queries returning different overlapping sets in order to ensure that all variances are returned for examination. In both exercises, the analysts report that all crimes that they would have wished to examine were contained within the clusters.

The results from Exercise 1 demonstrate that crimes within a single cluster have strong similarities and, as in the.results from cluster 2, may even contain crimes that have been committed by the same offender. With further refinements it should be possible to suggest names from the known offender list as being responsible for, as yet, unsolved crimes.

The results from Exercise 2 indicate that this type of model could be used as an initial match against the index case by restricting the search space to the Police Force area in which that crime occurred. A second pass through the data would include those crimes from the adjacent Force areas and a third pass could include national data.

This prototype system took ten weeks to develop from being unfamiliar with the data and its structures, to gain domain understanding, encode and model the data and pass the results through a verification process. Prior to and independent of this study, three persons; a medical psychologist, a statistician and a researcher, took two years to complete a study that reached broadly similar results. As a result of this study the NCF plan to commence an in depth 12 month pilot using the software.

7.1 Further Work

- Only two sets of two variable types were used in this study. There is scope to increase the number of sets and the number of variables within each set, model each and ascertain the behavioural consistency across each type. An example would be approach, verbal themes and precautions within the same model.
- Use several combinations of two variable sets and establish whether the same crimes are clustered together in more than one of the resulting models. The greater the number of crimes that are clustered together across the models may indicate that the same offender is responsible.

8. ACKNOWLEDGMENTS

Our thanks to the National Crime Faculty of the National Police College Bramshill United Kingdom for providing the data and independent verification of the results.

9. REFERENCES

[1] Adderley, R., Musgrove, P. B., (1999), *Data mining at the West Midlands Police: A study of bogus official burglaries*. BCS Special Group Expert Systems, ES99, London, Springer-Verlag, pp191-203

[2] Adderley, R.., Musgrove, P.B, (2001), Police crime recording and investigation systems, a user's view. *Policing An International Journal of Police Strategies and Management*, 24(1), pp100-114.

[3] Brantingham, P.L., Brantingham, P.L,. (1991), Notes on the geometry of crime, *in Environmental Criminology*, USA: Wavelend Press Inc.

[4] Canter, D., Heritage, R., (1990), A multivariate model of sexual offence behaviour: Developments in offender profiling I. *The Journal of Forensic Psychiatry*, 1(2), pp185-212.

[5] Canter, D., Larkin, P., (1993), Environmental range of serial rapists. *Journal of Environmental Psychology*, 13 pp63-69.

[6] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, Rudiger (2000) *CRISP-DM 1.0 Step-by-step data mining guide*, USA: SPSS Inc. CRISPWP-0800 2000.

[7] Davies, A., (1992), Rapists' behaviour: A three aspect model as a basis for analysis and the identification of serial crime. *Forensic Science International*, 55 pp173-194.

[8] Hazelwood, R. R., Reboussin, R., Warren, J. I. (1989), Series Rape: Correlates of increased aggression and the relationship of offender pleasure to victim resistance. *Journal of Interpersonal Violence* 4 pp65-78.

[9] Kohonen, T. (1984), Self-organisation and associative memory, *Springer series in information sciences*, Vol 8. Springer Verlag, New York, USA, 1984

[10] LeBeau, J.I., (1987), Patterns of stranger and serial rape offending: Factors distinguishing apprehended and at large offenders. *Journal of Criminal Law and Criminology* 78 pp309-326.

[11]Rhodes, W.M., Conly, C., (1991), *The criminal commute: A theoretical perspective in Environmental Criminology*, USA: Wavelend Press Inc.

[12] Weiss, S.M., Indurkhya, N., (1998), *Predictive data mining: a practical guide*. San Francisco, USA: Morgan Kaufman Publishers Inc.