International Research Journal of Computer Science (IRJCS)
Issue 7, Volume 2 (July 2015)
ISSN: 2393-9842
www.irjcs.com

A Proposed Data Mining Model to Enhance Counter-Criminal Systems with Application on National Security Crimes

Dr. Nevine Makram Labib
Department of Computer and Information Systems
Faculty of Management Sciences
Sadat Academy for Management Sciences, Egypt

Brigadier-General Wael Kamal Arafa
Department of Computer and Information Systems
Faculty of Management Sciences
Sadat Academy for Management Sciences, Egypt

Abstract— Data mining tools are becoming more powerful in different domains. These tools have proved efficient in the field of counter terrorism. In this study, some of the important data mining applications used in predictive policing are discussed. Examples will be drawn using real-world data collected from the Egyptian Ministry of Interior. The main purpose of this research is to propose and recommend a data-mining model that works best for predicting the most important factors affecting crime incidence. By using decision trees, Naïve Bayes, and Association rules while putting the prediction efficiency for each algorithm in comparison; the association rules came out with the best prediction probabilities.

Based on the results and conclusions drawn, it is recommended to implement the proposed model as a part of a working system that can be fed by police officers from transactional-level computer-systems directly from police stations to decision-making levels.

Keywords— Predictive Policing, Data mining, Decision Trees, Naïve Bayes, Association Rules

I. INTRODUCTION

Data mining (DM) was only used by large research institutions until tools were made easy for non- professionals to use. This led to applying DM techniques in different domains, including crime fighting. During the past decade, researchers focused on analysing crime records in order to uncover information and discover knowledge that might be the key for the puzzle. The process of predicting crime is also known as *Predictive Policing*. Many systems are currently in use by the law enforcement agencies in the United States (W. Perry et.al. 2014).

A. Problem Background:

Due to the increase of crime rate in Egypt; the Egyptian Ministry of Interior has been focusing on empowering Information Technology professionals and security researchers to develop and maintain information systems and patterns to help officers in their daily transactions. As for the higher management level, these systems can help in the decision making process. Currently developed, are systems made to help officers identify, import, display, and share valuable data.

B. Challenges:

One of the main challenges of developing information systems that deal with criminal records is the sensitivity of such data. According to the European Convention on Human Rights and the National Data Protection Authorities, criminals' data is considered private to their own. Moreover, the United Nations International Covenant on Civil and Political Rights (1966) stated that "No one shall be subjected to arbitrary or unlawful interference with his privacy". How is the collected data going to be handled, who will collect it, and who is going to be able to have access to it. Due to the sensitivity and privacy of criminal records; the major challenge obviously comes in getting the data to test.

Questions that arose were how to get unclassified data, whether it is possible to scrub and clean the classified data and produce reasonable data at the unclassified level. How can large datasets consisting of multimedia data types be found, and whether it is possible to develop training datasets where one can apply the various DM tools to determine their efficiency, that is, the possibility of making use of the given data.

C. Aim of the Research:

The aim of this research is to propose a DM model that is suitable to be applied on criminal records for predicting the most important factors affecting crime incidence.

D. Research Scope and Framework:

Criminal records data from 1996 to 2012 related to crimes that happened in Alexandria, Egypt were collected identified, and criminals' profiles were prepared. These profiles consist of criminals' personal information including age, profession, mental and educational level, social class, geographic locations of their past crime(s), and types of committed crimes.

A database was created including criminals records. It was used as a basis for the machine learning and DM tools that will be reviewed in the research.

Recommendations are to be given according to the results of testing various mining techniques on the collected set of records.

II. LITERATURE REVIEW OF DATA MINING AND COUNTER-TERRORISM

Data mining is the process of posing queries and extracting useful patterns or trends often previously unknown from large amounts of data using various techniques such as those from pattern recognition and machine learning (Thuraisingham, B., 2003).

Two main forms of data analysis are normally used to extract models describing important classes or predict future data trends: Predictive and descriptive forms.

Regarding descriptive forms, they are commonly referred to as Classification. This data analysis helps providing a better understanding of large data. Classification predicts categorical and prediction models predict continuous valued functions. As for regression analysis, it is a statistical methodology that is most often used for numeric prediction.

The RAND research on predictive policing (2013) classifies prediction methods into four main categories as follows:

methods for predicting crimes, methods for predicting offenders, methods for predicting perpetrators' identities, and methods for predicting victims of crimes.

John Eck, et.al. (2005) in their book "Mapping Crime" have proved how effective graphical tools can be in extracting and displaying critical notes; hotspot analysis and crime mapping are two good examples for graphing. Since crime is assumed to happen in the same place repeatedly, hotspot analysis makes a geographical mapping to find "hotspots" where crimes would occur. Usually, in countries with high crime rate, finding hotspots will not be that important. For this reason, hotspot analysis uses many techniques, and graphs are usually drawn as levels where each level gives a different message. The following figure explains the notes taken by a given level of hotspot graph.

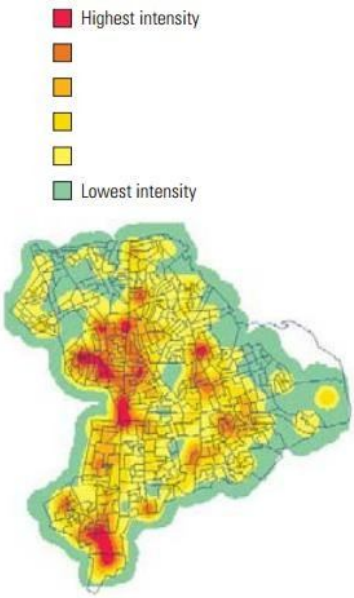

Fig. 1 Quartic kernel density estimation surface for vehicle crime using a bandwidth of220m.Source: Predictive policing: The role of crime forecasting in law enforcement operations [2].

Interpolation is one of the most commonly used methods for visualizing the distribution of crime and identifying hot spots. It aggregates points within a specified search radius and creates a smooth, continuous surface that represents the density or volume of crime events distributed across the area as displayed in the picture above (McCue, et. al., 2003).

Moreover, Chen, H., et. al. (2003) were motivated by the concern about national security that arose after the 9/11 attacks. They depended on COPLINK case studies to implement DM patterns in order to prove the effectiveness of DM in crime prediction.

III. DATA PREPARATION AND TRANSFORMATION

In this research, data was collected from the Egyptian Ministry of Interior records. The main dataset was created using MS SQL Server 2008. Then, MS SQL Server Business Intelligence Development Studio was used to implement the mining algorithms and produce the visual results.

In this paper, various features were implemented to support various tasks that can be performed on a given dataset. These features include:

- Filtering is an important task as some DM tasks such as clustering (DBScan, EM) and Association (Apriori) involving large amounts of data takes up large amount of time and memory. Not performing this step will slow down the application, to counter this problem the initial dataset instances which are populated by all the database attributes are filtered down to the most useful attributes depending on the user's interest and requirements.

- Classifiers: Nominal and numeric attributes are to be predicted using implemented algorithms such as Naïve Bayes and Decision Trees.
- Association: is a gigantic move towards the success of DM in improving crime prediction. It finds relations between crime records attributes as crime type, and place of crime incidence.
- Clustering normally finds groups in the dataset with similar instances.
- Data is fetched from different sources. A new instance is created to store the imported dataset. An instance is an object that stores all the attributes and data objects of the dataset. Functions such as filter, clustering, and association, are performed on such instances. This can be achieved by using SQL statements to fetch the data based on the dataset attribute. After the datasets are fetched into the application, different DM tasks are implemented.

A. Creating Datasets:

1) Data Selection: A total of 351 cases were collected. The created database depended mainly on reviewing and summarizing the collected cases in order to find helpful information.
2) Attributes were carefully selected. Some of the records were missing one or more values. Moreover, some attributes as "the motive for the crime" was found missing in many cases, although it would have been very helpful to analyze and include the motive in the set of attributes.
3) Data Cleaning included removing duplicates and dealing with missing values.
4) Data Reduction was a crucial step in order to set the database in a well-ordered summarized form that can be used by mining tools. Reduction was made by using normalization, and aggregation.
5) Discretization was applied to categorize data as numeric intervals, since not all DM algorithms can work with numeric values.

B. Practical Implementation of Algorithms:

1) Building Data Structure: The data was discretized and grouped to be inserted into the database as a set of attributes. The following table shows the codes of crime types and the percentage of their occurrence in the collected cases.

TABLE I
PERCENTAGES OF CRIME TYPES IN THE DATASET. CODES SHOWN IN APPENDIX

Crime Type	Percentage
P	64.25%
Q	29.25%
R	6.23%
Missing Values	0%

2) Implementing Decision Trees: As shown in the figure 2, level 1 describes the set of all cases that is split into two nodes i.e. Education (whether educated or not). Decision trees showed the importance of the dependency between crimes attributes. In the context of all cases, the darker the attribute gets in the figure, the more effective it is in crime occurrence. It is clear in the shown tree that most drug dealers are uneducated males who have no profession.

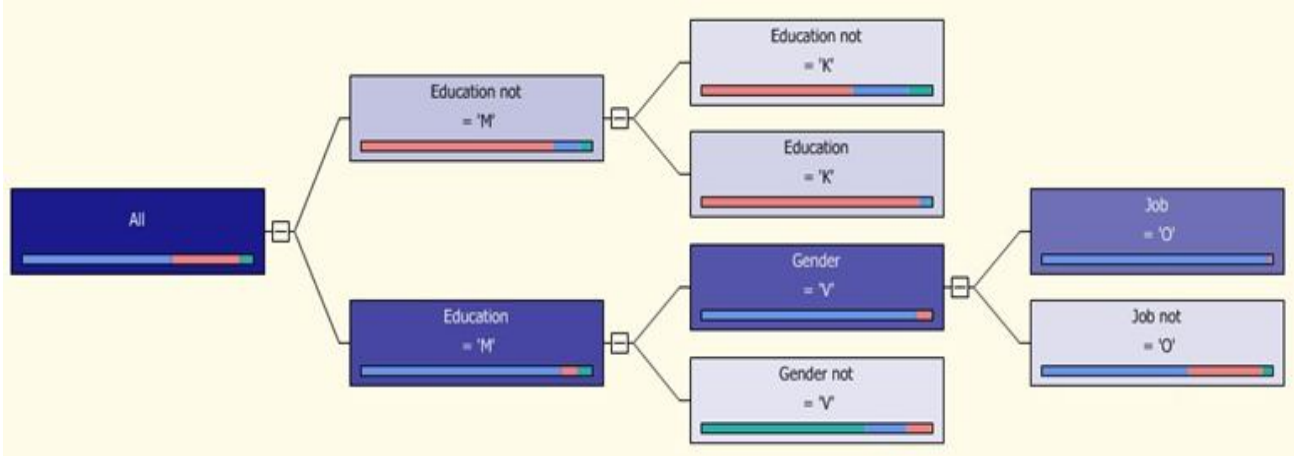

Fig. 2 Microsoft Decision Tree-MS SQL Server Business Intelligence Development Studio

3) Implementing Naïve Bayes:

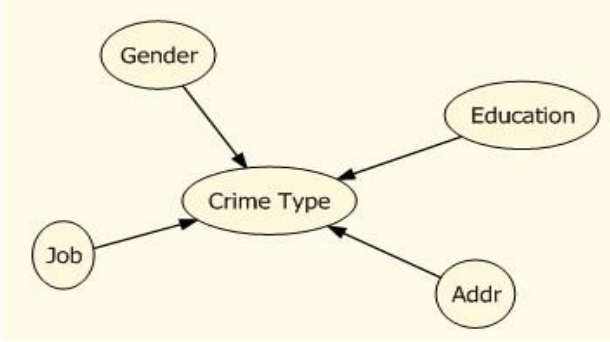

Fig. 3 Naïve Bayes Dependency Network

The Naïve Bayes dependency network describes the most important attributes affecting crime occurrence are Address, Education, Gender, and Job (sorted alphabetically). While after reviewing the attributes characteristics shown in figure3 the importance of the attributes towards drug dealing crime prediction as an example are as follows:

TABLE II
THE IMPORTANCE OF ATTRIBUTES AFFECTING DRUG DEALING CRIME PREDICTION

Attribute	Value	Percentage
Gender	V	97.5%
Education	M	93.3%
Job	O	87.2%
Address	F	40.6%

We can conclude from the resulting data that -by taking percentages average- 79.6% of drug dealers are uneducated males who have no occupations and living in Bedouin environments.

4) Implementing Association Rules:

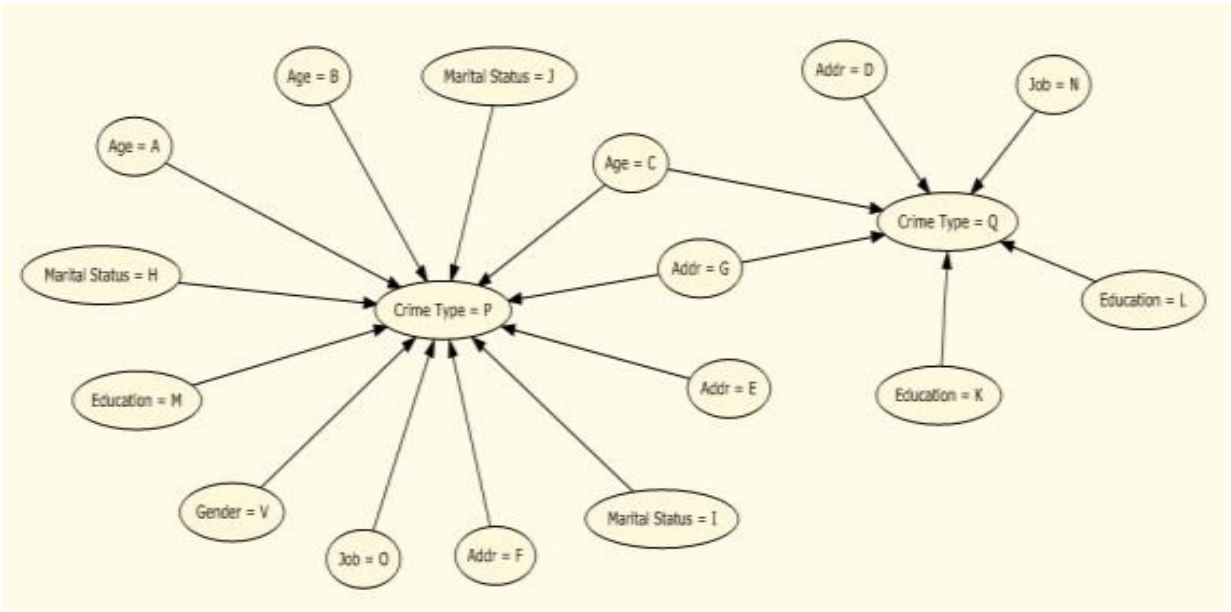

Fig. 4 Dependency Network – MS Association Model – MS SQL Server Business Intelligence Studio

Not all of association rules may be suitable for prediction. L. Jiang (2005) introduced an approach called "predictive mining" that discovers a set of prediction rules given a suitable set of data like the one we are working on. MS Association was implemented on the same data set to give the following outcomes:

Fig. 5 Attribute Characteristics for P – Naïve Bayes Algorithm – SQL Server Business Intelligence Development Studio

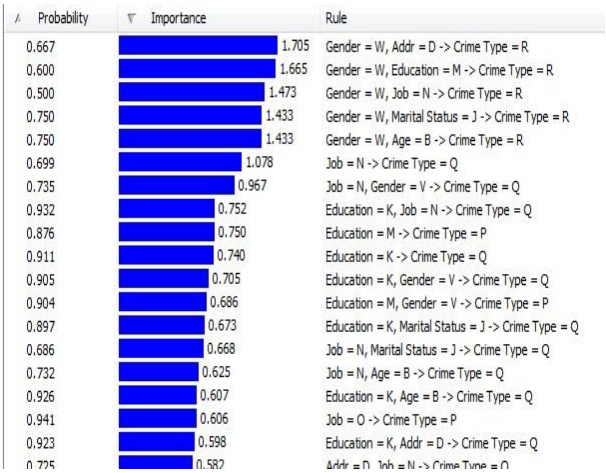

Fig. 6 Association Rules –Showing 19/140 rules

5) Evaluating Models Efficiency:

The efficiency of` the proposed models were evaluated using the Lift Chart to compare the prediction probability in the implemented mining models:

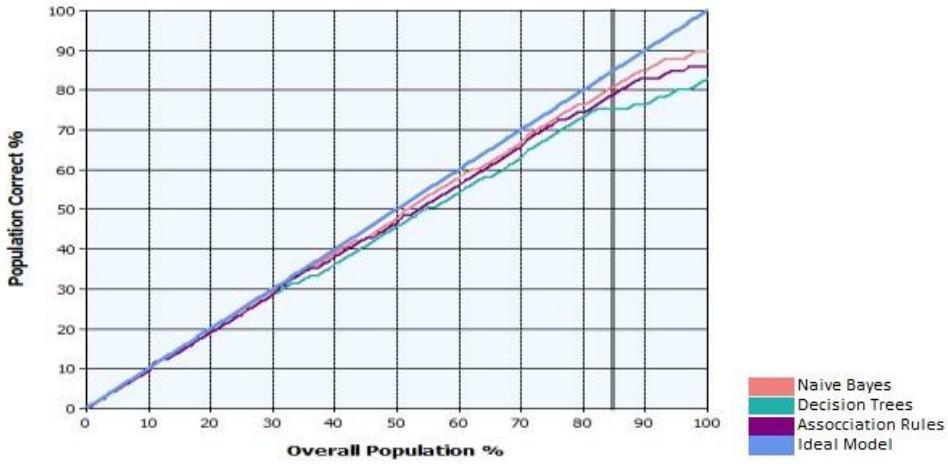

Fig. 7 Lift Chart for the three prediction models grouped. MS SQL Server Business Intelligence Studio

The line that gets closer to the "ideal model" line gets an overall higher probability rank.

IV. FINDINGS

- The decision tree results gave us a solid proof that education is the most crucial factor affecting crime occurrence. Gender comes in the next level, then the job.
- Naïve Bayes gave a clear recognition of which attribute makes the highest effect on the prediction process.
- Association rules can be filtered to show the most effective factors on a certain rule. For example, when we add a "Crime Type = P" filter, the resulting rules will be:

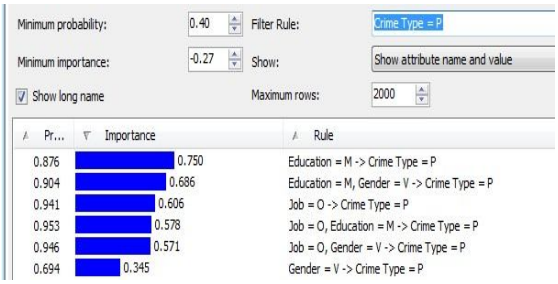

Fig. 8 Filtered Association Rules – MS SQL Server Business Intelligence Development Studio

The results support the fact that education is the main factor affecting crime incidence.

After evaluating the models' efficiency, the Lift Chart clearly showed how the Association rules model line kept closer distance to the ideal line than both Naïve Bayes and decision tree models. More accurately, the results of the prediction probability came out as follows:

TABLE III
COMPARISON BETWEEN THE PREDICTION PROBABILITIES FOR TESTED MINING ALGORITHMS

Model	Prediction Probability
Association Rules	98.11%
Naïve Bayes	97.81%
Decision Tree	92.07%

V. CONCLUSIONS

The main purpose of this research was to propose and recommend a DM model that works best for predicting the most important factors affecting crime incidence. By using decision trees, Naïve Bayes, and Association rules while putting the prediction efficiency for each algorithm in comparison; the association rules came out with the best prediction probabilities.

ACKNOWLEDGMENT

The research was supported by the General Administration of Technical Assistance in the Egyptian Ministry of Interior, although they may not agree with all of the conclusions of this paper.

REFERENCES

[1] W. Perry, B. McInnis, C. Price, S. Smith, and J. Hollywood (2013) "Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations. RAND Corporation" on RAND. [Online]. Available: http://www.rand.org/pubs/research_reports/RR233.html
[2] J. Eck, S. Chainey, J. Cameron, M. Leitner, and R. Wilson (2005) "Mapping Crime: Understanding Hot Spot" on London's Global University website. [Online]. Available: http://eprints.ucl.ac.uk/11291/1/11291.pdf
[3] C. McCue, and A. Parker, "Connecting the Dots: Data Mining and Predictive Analytics in Law Enforcement and Intelligence Analysis," in *The Police Chief*, 2003.
[4] B. Thuraisingham (2003). *Web Data Mining and Applications in Business Intelligence and Counter-Terrorism.* Ch.3.
[5] (2014) "Handbook on European Data Protection Law" on European Court of Human Rights homepage. [Online]. Available: http://www.echr.coe.int/Documents/Handbook_data_protection_ENG.pdf
[6] F. Cate, "Harvard Civil Rights-Civil Liberties Law Review (CR-CL)," 2008, Vol. 43, No.2.
[7] R. Okonkwo, and F. Enem (2011). ITePED2011 homepage of NCS. [Online]. Available: http://www.ncs.org.ng/wp- content/uploads/2011/08/ITePED2011-paper10.pdf
[8] O. Fredrick, and O. Patrick, "Survey of Data Mining Methods for Crime Analysis and Visualization," in Advances in System Modelling and ICT Applications, 2006, vol.2.
[9] R. Popp, and J. Poindexter, "Countering Terrorism through Information and Privacy Protection Technologies" in IEEE Computer Security, vol.4, p. 6
[10] J. Deogun, and L. Jiang, "Prediction Mining – An Approach to Mining Association Rules for Prediction" in Lecture Notes in Computer Science, 2005, vol. 3642, pp. 98-108.
[11] H. Chen, W. Chung, Yi Qin, M. Chau, J. Jie Xu, G. Wang, R. Zheng, and H. Atabakhsh, "Crime Data Mining: An Overview and Case Studies", in Proceedings National Conference on Digital Government Research, 2003, pp. 50-51.
[12] A. Malathi, S. Santhosh, Baboo, and A. Anbarasi, 2011. "An intelligent Analysis of a City Crime Data Using Data Mining". International Conference on Information and Electronics Engineering, Vol. 6. CONFERENCE PAPAER

APPENDIX

The following table contains the codes used in the data set for each attribute and their corresponding values:

Age		Educational Level	
A	15-30 years	K	Higher education
B	30-50 years	L	Intermediate
C	50-above	M	None
Gender		**Job**	
V	Male	N	Employed
W	Female	O	Unemployed
Address		**Crime Type**	
D	Civilized area	P	Drug Dealing
E	Uncivilized area	Q	Public Funds
F	Bedouin area	R	Public Morals
G	Rural area		
Marital Status		**Judgment**	
H	Single	S	1-7 years in prison
I	Married	T	7-15 years in prison
J	Married with children	U	15 years-above